BELIEF - Bayesian Sign Entropy Regularization for LIME Framework

Revoti Prasad Bora[1] **Philipp Terhörst**[2] **Raymond Veldhuis**[1] **Raghavendra Ramachandra**[1] **Kiran Raja**[1]

[1]Norwegian University of Science and Technology, Gjøvik, Norway
[2]Paderborn University, Paderborn, Germany

Abstract

Explanations of Local Interpretable Model-agnostic Explanations (LIME) are often inconsistent across different runs making them unreliable for eXplainable AI (XAI). The inconsistency stems from sign flips and variability in ranks of the segments for each different run. We propose a Bayesian Regularization approach to reduce sign flips, which in turn stabilizes feature rankings and ensures significantly higher consistency in explanations. The proposed approach enforces sparsity by incorporating a Sign Entropy prior on the coefficient distribution and dynamically eliminates features during optimization. Our results demonstrate that the explanations from the proposed method exhibit significantly better consistency and fidelity than LIME (and its earlier variants). Further, our approach exhibits comparable consistency and fidelity with a significantly lower execution time than the latest LIME variant, i.e., SLICE.

1 INTRODUCTION

Explanation methods for Deep Learning (DL) models need access to intermediate layers which often is not easy to access as in the case of Class Activation Map (CAM) based methods (e.g., Grad-CAM [Selvaraju et al., 2017], Grad-CAM++ [Chattopadhay et al., 2018]). Addressing such a problem, model agnostic based explanations methods have been proposed which do not need access to the layers of DL models and work in complete black-box setting. Methods like LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017] and their variants have shown the application in black-box setting making them a popular choice in post-hoc explanations.

Despite the popularity and model agnostic property of LIME [Ribeiro et al., 2016], a number of inconsistency in LIME has been reported [Zhang et al., 2019, Gosiewska and Biecek, 2019, Li et al., 2023, Lee and Lee, 2023, Zhao et al., 2021, Zafar and Khan, 2019, Zhou et al., 2021]. Gosiewska and Biecek [2019] and Lee and Lee [2023] highlight the instability of additive explanations and observed variations in feature importance across different methods. Additionally, Zhang et al. [2019] note (i) variability in explanations due to sampling, (ii) dependence on hyper-parameters such as neighborhood size and sample count, and (iii) fluctuations in model reliability across different instances. These factors lead to inconsistency in explanations making them unreliable.

Figure 1: Figure showing the top five positive and negative superpixels (segments) of inconsistent LIME explanations for a random image of the Oxford-IIIT Pets dataset with Inception V3 model for four different runs. The predicted class was Newfoundland, and the prediction probability was 0.46. Blue and red colors denote positive and negative superpixels, and the numbers inside the superpixels specify their importance rank. By addressing the limitations, we demonstrate consistent explanation using our proposed approach for the same image and model. (Results in Figure S1 - supplementary material).

As shown in Figure 1, the inconsistency of LIME explanation can be noted in highlighted superpixels that flip between positive (blue) and negative (red) contributions for the output probability. Further, it can be noted that the importance ranks of the superpixels (segments) for both positively and negatively contributing superpixels also vary across different runs. These inconsistencies make interpretability challenging [Bora et al., 2024]. This flipping of superpixel sign, for different independent runs, is defined as the uncer-

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:332–354.

tainty in the sign of superpixels (i.e., Sign Entropy [Bora et al., 2024]). Estimating the uncertainty of the signs of the superpixels by using bootstrapping on frequentist Ridge Regression, and eliminating superpixels with high uncertainty in signs (i.e., selecting features with low sign entropy) has been shown to stabilize LIME explanations [Bora et al., 2024]. This however, comes at the cost of considerable increase in execution time due to bootstrapping approach. In this paper, we propose a novel Sign Entropy Regularization using Bayesian paradigm to estimate the uncertainty and mitigate the inconsistencies while achieving significantly faster ($\approx 10\times$) execution time.

2 RELATED WORKS

Several works have studied different approaches to mitigate the inconsistency of LIME. ALIME [Shankaranarayana and Runje, 2019] incorporates an autoencoder as a weighting mechanism to assess the proximity of sampled points to the instance being explained (IE), thereby improving coefficient stability. DLIME [Zafar and Khan, 2019] applies hierarchical clustering to segment the dataset into multiple clusters and selects representative points from the cluster nearest to the IE, ensuring alignment with LIME's locality principle. S-LIME [Zhou et al., 2021] enhances the consistency of LIME explanations by introducing a hypothesis testing framework, which utilizes the Central Limit Theorem (CLT) to determine the required sample size for stable explanations. BayLIME [Zhao et al., 2021] adopts a Bayesian approach for local surrogate modeling, where explanations are generated by combining prior knowledge with estimates from newly sampled data through a Bayesian-weighted summation. SLICE [Bora et al., 2024] uses a two stage strategy of using adaptive Gaussian Blur (Adaptive-blur) followed by feature selection algorithm to remove inconsistent superpixels from explanations. The feature selection algorithm uses bootstrapping to estimate the probability of sign flips of coefficients (i.e., sign entropy) and eliminate superpixels with a high likelihood of sign flips. A surrogate model is built using Ridge Regression, similar to the original LIME [Ribeiro et al., 2016]. SLICE [Bora et al., 2024] achieved high consistency and fidelity as compared to LIME [Ribeiro et al., 2016] and BayLIME [Zhao et al., 2021] at the cost of a considerable increase in execution time due to the use of bootstrapping for feature selection.

3 OUR CONTRIBUTIONS

We propose a new method, BELIEF, to achieve consistency of explanations similar to SLICE [Bora et al., 2024], using a Bayesian Regularization approach, eliminating the need for external feature selection and offering faster computation time. Specifically, we propose a novel Sign Entropy regularization, modeling it in Bayesian paradigm instead of the bootstrapping-based frequentist approach. Thus, our approach achieves consistency and fidelity similar to SLICE but with much faster computing time ($\approx 10\times$) by eliminating bootstrapping based feature selection.

We first demonstrate our proposed method for reducing the uncertainty associated with the sign of the coefficients in linear models, thereby enhancing consistency on tabular datasets. We then show the broader use of our proposed method in stabilizing LIME for consistent explanations for DL-based image classification applications. With extensive experiments on multiple tabular and image datasets, we evaluate and compare our approach with State-Of-The-Art (SOTA) counterparts for consistent explanation. Additionally, we perform statistical validation using multiple tests to provide empirical evidence supporting our claims of consistent explanation, high fidelity, and lower execution time.

4 PROPOSED APPROACH

LIME works by building a local simple surrogate model to approximate the decision boundary near the IE (details in Appendix A.1). A frequentist Ridge Regression is used as a surrogate model in the LIME and SLICE implementation, while in BayLIME, a Bayesian Ridge Regression is used. The coefficients of the surrogate model, which have mapping to each superpixel, represent the impact (i.e., sign and magnitude) of the corresponding superpixels on the output probability. Hence, the flipping of the sign of the surrogate model coefficients leads to uncertainty regarding the direction of impact of the superpixels on the output probability. This section discusses our approach to reducing the uncertainty of coefficients' signs in the surrogate model using our novel Sign Entropy regularization. Additionally, the relative ranks of the coefficients also stabilize as an added advantage of our regularization, further enhancing explainability. We first discuss it as a general-purpose regularization technique, and then in subsequent sections, we show its applicability on tabular and image datasets.

4.1 BAYESIAN FORMULATION

Bayesian Ridge Regression model is defined as $y = X\beta + \epsilon$ with β representing the vector of coefficients, and ϵ representing Gaussian noise with precision parameter α. Bayesian Ridge Regression applies a prior $p(\beta \mid \lambda)$ over β with Gaussian distribution $\mathcal{N}$ given by:

$$p(\beta \mid \lambda) = \mathcal{N}(0, \lambda^{-1} I),$$

where, λ controls the regularization strength i.e., the precision of the prior and I is an identity matrix.

The likelihood function for the observed data y, given X

and β is Gaussian:

$$p(y \mid X, \beta, \alpha) = \mathcal{N}(y \mid X\beta, \alpha^{-1}I),$$

where, α represents the precision of the noise.

Using Bayes' theorem, the posterior distribution over β, given X and y, is obtained as:

$$p(\beta \mid X, y, \alpha, \lambda) = \mathcal{N}(\beta \mid \mu_\beta, \Sigma_\beta),$$

where, mean μ_β and covariance Σ_β are given by:

$$\mu_\beta = \alpha \Sigma_\beta X^T y \quad \Sigma_\beta = (\alpha X^T X + \lambda I)^{-1},$$

λ and α are the hyper-parameters of Bayesian Ridge Regression normally with a γ distribution prior.

Bayesian Ridge Regression follows an iterative Bayesian update process, where the posterior at each step serves as the prior for the next iteration [Tipping, 2001], [MacKay, 1992]. We extend this approach by enforcing a Sign Entropy prior dynamically during optimization. Instead of using a Gaussian prior that does not enforce sign stability, we enforce the Sign Entropy prior to refine the feature set at each iteration based on the posterior distribution of the coefficients. This ensures that only stable features contribute to learning in subsequent iterations. The sparsity enforcing Sign Entropy prior in our approach acts as a structured regularization mechanism for capturing stable/consistent coefficient estimates. The proposed Sign Entropy Regularization is further discussed in detail in the next sub-section.

4.2 SIGN ENTROPY REGULARIZATION

For a given j^{th} coefficient β_j, the variance σ_j^2 is given by:

$$\sigma_j^2 = \Sigma_\beta[j, j]$$

As the posterior distribution of β_j is $\mathcal{N}(\beta_j \mid \mu_\beta, \Sigma_\beta)$, we can calculate the probability that β_j is positive (p^+) as follows:

$$\begin{aligned} p^+ &= P(\beta_j > 0) = 1 - P(\beta_j \leq 0) \\ &= 1 - \Phi\left(-\frac{\mu_j}{\sigma_j}\right) \end{aligned}$$

where, μ_j is the posterior mean and σ_j is the variance of of β_j, and Φ is the Cumulative Distribution Function (CDF) of the standard normal distribution.

The Sign Entropy $H(\beta_j)$ is computed using p^+ and p^- (i.e., $p^- = 1 - p^+$) as below:

$$H(\beta_j) = -p^+ \log_2(p^+) - p^- \log_2(p^-),$$

where, p^+ is the estimated probability that β_j is positive and $p^- = 1 - p^+$ is the estimated probability that β_j is negative. A high value of Sign Entropy indicates that the coefficient's sign has a high probability of flipping.

The Sign Entropy prior applied on the coefficients at each iteration enforces sparsity by eliminating features with high entropy:

$$\mathcal{F}^{(t+1)} = \mathcal{F}^{(t)} \setminus \{j \mid H(\beta_j) > \zeta\}$$

where, $\mathcal{F}^{(t)}$ represents the set of active features in a particular iteration t, $\setminus$ denotes set minus, and features with high Sign Entropy $H(\beta_j) > \zeta$ are eliminated from the model in the next iteration of the optimization process, and ζ is a hyper-parameter representing the highest acceptable threshold for Sign Entropy (details in Appendix B).

4.3 EVALUATION OF SIGN ENTROPY REGULARIZATION

We first validate the proposed Sign Entropy regularization by comparing it with a family of other approaches with different regularization strategies. We compared our approached with frequentist Lasso [Tibshirani, 1996], Ridge [Hoerl and Kennard, 1970], Bayesian Ridge [MacKay, 1992] Tipping [2001], Automatic Relevance Determination (ARD) [MacKay, 1992], [Salakhutdinov, 2024] and Ordinary Least Squares (OLS) [Kutner et al., 2005]. We compute Average Sign Flip Entropy (ASFE) [Bora et al., 2024] and Root Mean Square Error (RMSE) for all approaches to establish the efficacy of our newly proposed Sign Entropy-based regularization on two public datasets: House Prices - Advanced Regression Techniques dataset from Kaggle [Kaggle, 2024] and Appliance Energy Prediction dataset [Candanedo et al., 2017] from the UCI repository. We used the implementation from Scikit-learn [Pedregosa et al., 2011] for the SOTA approaches and we wrote our code[1] in python.

We evaluate LASSO, Ridge, Bayesian Ridge, and the proposed method with different settings of the regularization hyper-parameter ($\alpha = 0.1, 0.5, 1$ in case of LASSO and Ridge and $\lambda_{init} = 0.1, 0.5, 1$ for Bayesian Ridge)[2]. This enabled us to compare the proposed regularization method with other methods at different regularization strengths. We then computed the ASFE and the RMSE[3] metrics by performing five-fold cross validation with five repeats. As noted from Figure 2, the proposed regularization scheme achieves low ASFE compared to all other approaches indicating stability/consistency of coefficient sign. Further, we note no significant loss of RMSE as compared to other approaches.

[1] `https://github.com/rebathip/BELIEF.git`

[2] OLS does not have a regularization term and ARD does not have λ_{init} hyper-parameter.

[3] We normalized RMSE to a scale of [0,1] using min-max scaling.

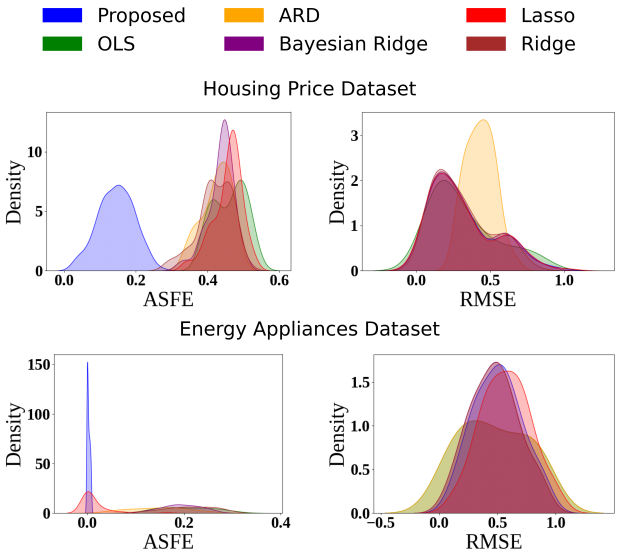


Figure 2: Distribution of ASFE and RMSE scores of the proposed method and other methods for Housing Price and Energy appliances datasets. The proposed method achieves much lower ASFE score while maintaining comparable RMSE with other methods.

The ASFE score for the proposed method outperformed other methods by a large margin which can be observed from the low overlap of the ASFE scores. For ascertaining that our method does not impact the predictive power, we conducted Two-sample Kolmogorov-Smirnov (KS) test [Hodges Jr, 1958] on the distribution of RMSE scores of our proposed method with the other methods. We used KS test owing to its non-parametric nature. Null Hypothesis H_0 was that the two distributions are identical and the alternate hypothesis was that they are not identical. The p-values (refer to supplementary for details Table S9) from the tests were much higher than the commonly accepted threshold of 0.05 providing insufficient statistical evidence to reject the Null Hypothesis H_0. Thus, we see that our method achieves high stability in terms of coefficients' sign flips while retaining comparable predictive power.

5 ON THE APPLICATION TO IMAGE EXPLANATIONS

We use the proposed Sign Entropy regularization method to provide consistent explanations for images. Additionally, to generate perturbed images for learning the surrogate model, we used Adaptive-blur from [Bora et al., 2024]. We conduct a series of experiments to demonstrate the applicability of the proposed approach to obtain consistent explanations. Two pre-trained image classification models - InceptionV3 [Szegedy et al., 2016] and ResNet50 [He et al., 2016] initialized with ImageNet weights on the Oxford-IIIT Pet Dataset Parkhi et al. [2012] and Pascal VOC 2007 [Everingham et al., 2007] dataset were used to evaluate the proposed approach[4].

We compare our method against LIME, BayLIME and SLICE. We use BayLIME with Grad-CAM as prior for our comparison, as it was demonstrated to have superior consistency and fidelity in comparison to LIME [Zhao et al., 2021]. We randomly selected 50 images from each of the mentioned datasets and analyze both DL models for 20 repeated and distinct runs. We computed the consistency and fidelity scores for each image-model and averaged them across all the 20 distinct runs. We follow similar settings as outlined in [Bora et al., 2024] to conduct our ablation study as presented in Table 1. Statistical significance of our findings is provided using Wilcoxon Signed Rank test [Virtanen et al., 2020] benefiting from a non-parametric nature of the test. The threshold of p-value to reject the Null Hypothesis is set at the commonly used threshold of 0.05 and to measure the effect size, we have employed Common Language Effect Size (CLES) [McGraw and Wong, 1992] [Vargha and Delaney, 2000].

5.1 EVALUATION METRICS

For a fair comparison of consistency and fidelity of our proposed approach with the SOTA approaches we use the Combined Consistency Metric (CCM) from [Bora et al., 2024]. CCM is defined as below:

$$CCM^{xp}_{M,I} = (1 - ASFE^{xp}_{M,I}) * ARS^{xp}_{M,I}$$

where, $ASFE^{xp}_{M,I}$ denotes the Average Sign Flip Entropy of the coefficients and $ARS^{xp}_{M,I}$ denotes the Rank Similarity of the superpixels in the explanations for a model M and image I. $ASFE^{xp}_{M,I}$ ranges from 0 to 1 with a lower value indicating more consistency and $ARS^{xp}_{M,I}$ ranges from 0 to 1 with a higher value indicating better consistency. $CCM^{xp}_{M,I}$ ranges between [0,1] where 0 denotes low consistency and 1 denotes full consistency in both Sign Entropy and superpixel importance ranks (Details provided in Appendix A in supplementary). Further, adapted Area Under Perturbation Curve (AOPC) [Bora et al., 2024] and Insertion and Deletion Area Under the Curve metrics [Petsiuk et al., 2018] are used for measuring the fidelity of explanations.

6 CONSISTENCY OF BELIEF

The Empirical Cumulative Distribution Function (ECDF) plots of the CCM scores for both models with Oxford-IIIT Pets dataset are provided in Figure 3 (refer Figure S3, and

[4]We use a kernel size of (5,5) for Gaussian Blur (similar to Bora et al. [2024]). The code for SOTA methods were obtained from the github repositories: BayLIME [Zhao, 2023], LIME and SLICE [Bora et al., 2024].

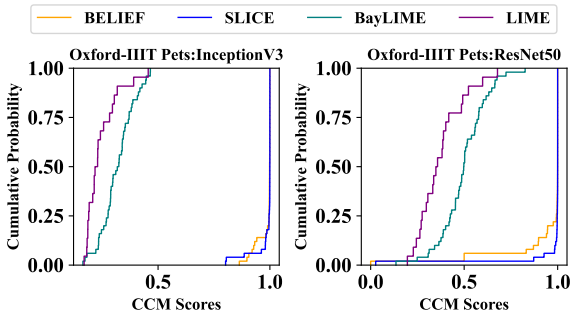


Figure 3: ECDF plot of CCM Scores for BELIEF, LIME, BayLIME and SLICE (higher score is better)

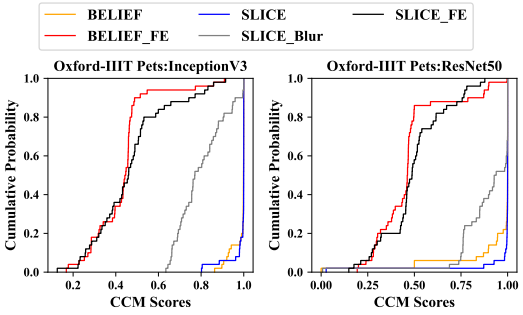


Figure 4: ECDF plot of CCM Scores for BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE and SLICE (higher is better)

Figure S5 in supplementary for exhaustive ECDF and Density plots). BayLIME and LIME have much lower CCM scores than BELIEF and SLICE. The results empirically indicate that our proposed approach performs at par, in terms of consistency, with SLICE without the additional step of feature selection.

Further, we conducted the Wilcoxon signed-rank test to ascertain that the higher CCM scores of BELIEF as compared to LIME and BayLIME are statistically significant. The p-values from the Wilcoxon Signed Rank tests were low (in the range of 8.9e-16 to 2.2e-11), the Test Statistics were high (in the range of 1227 to 1275) and effect sizes were large (in the range of 0.96 to 1) (refer Table S2 in supplementary for test details). The notably low p-value and the substantially high value of the Test Statistic provide robust statistical evidence to reject the null hypothesis. Further, the large effect sizes indicate that the higher CCM scores of BELIEF explanations were not only statistically significant but also practically meaningful.

Table 1: Ablation settings with BELIEF and SLICE variants

Method	Feature Elimination	Adaptive-Blur
SLICE_blur	✗	✓
SLICE_FE	✓	✗
SLICE	✓	✓
BELIEF	✓	✓
BELIEF_FE	✓	✗

6.1 ABLATION STUDY FOR BELIEF

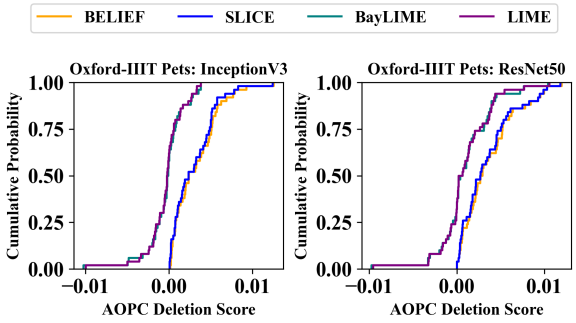
In our ablation study, we evaluate BELIEF in settings similar to SLICE, i.e., with (BELIEF) and without (BELIEF_FE) adaptive blur as noted in Table 1. As our proposed approach enforces sparsity using Sign Entropy regularization, BELIEF does not have a counterpart for SLICE_blur for ablation studies. The ECDF plots of the CCM scores of all the five variants of BELIEF and SLICE on Oxford-IIIT pets dataset are presented in Figure 4 (refer Figure S4 in supplementary for exhaustive plots). With similar ECDF plots, it can be seen that BELIEF and SLICE performed the best followed by SLICE_blur, while BELIEF_FE and SLICE_FE performed the worst. This supports our idea that using Sign Entropy as a regularization technique in the Bayesian paradigm can achieve the same consistency as SLICE without the need for an additional feature selection step.

Similarly, BELIEF_FE and SLICE_FE have higher CCM scores than LIME but lower than that of BELIEF, SLICE and SLICE_blur as seen in Figure 3, and Figure 4 (refer to exhaustive ECDF plots in Figure S4 and density plots in Figure S6 of supplementary). Without Adaptive-Blur, the created perturbed images, used in building the surrogate model, are significantly different from the original image making it difficult for BELIEF_FE and SLICE_FE to estimate the Sign Entropy of the superpixels/segments. However, when we combine Adaptive-blur with Sign Entropy regularization (or Sign Entropy based feature selection in SLICE) the estimation of Sign Entropy is more accurate leading to the proper elimination of inconsistent features/superpixels. We further performed Wilcoxon Signed Rank tests to ascertain the statistical significance of our claims. The low p-values (8.9e-16 to 4.7e-3) and high value of Test Statistics (904 to 1275) provide robust statistical evidence that the CCM scores of BELIEF is higher than LIME, BayLIME, SLICE_blur and SLICE_FE. Further, the effect size (close to 1 in most cases) support that our observations are statistically significant and practically meaningful. The details of the tests are in Table S3 of the supplementary material.

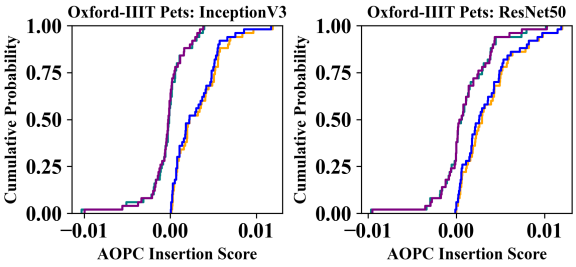
7 FIDELITY EVALUATION OF BELIEF EXPLANATIONS

7.1 AREA UNDER PERTURBATION CURVE (AOPC)

The AOPC scores for BELIEF was higher than that of LIME and BayLIME as seen in Figure 5 for Oxford-IIIT Pets

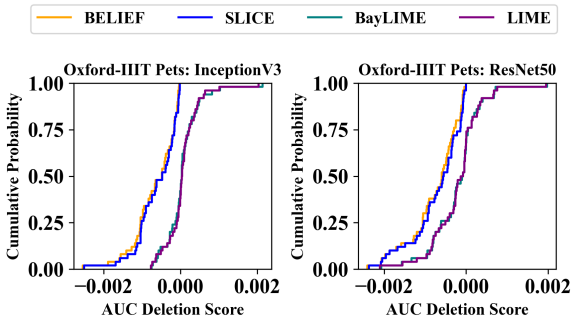


(a) ECDF plots of AOPC deletion scores

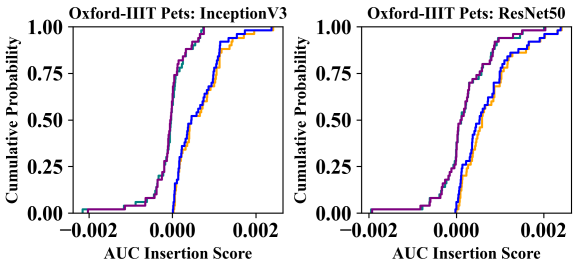


(b) ECDF plots of AOPC insertion scores

Figure 5: ECDF plots of AOPC (Higher AOPC indicates higher fidelity)



(a) ECDF plots of AUC deletion scores (Lower is better)



(b) ECDF plots of AUC insertion scores (Higher is better)

Figure 6: ECDF plots of AUC scores for Oxford-IIIT Pets Dataset for Inception V3 and ResNet50

Dataset (Refer Figure S7a and Figure S7b in Supplementary for both datasets). The AOPC scores are low as it is based on the difference between the output probability of the unperturbed image and the perturbed images. We performed Wilcoxon Signed Rank tests along with effect size calculation to ascertain that the higher AOPC scores of BELIEF explanations, as compared to those of LIME and BayLIME, were statistically significant and practically meaningful. In our tests, as shown in Table 2 for Oxford-IIIT Pets dataset, the p-values were low, the test statistics were high and the effect sizes were close to 1. The details of test results for both the datasets are provided in Table S5 of supplementary material. These provide robust statistical evidence confirming that the AOPC scores of BELIEF were significantly higher than those of LIME and BayLIME and are practically meaningful.

Table 2: Wilcoxon signed rank test results for comparison of LIME, BayLIME, and BELIEF. For a given pair (x,y), the null hypothesis H_0 was "The median of the differences $(metric(x) - metric(y))$ is equal to zero," and the alternative hypothesis was H_a was "The median of the differences $(metricscore(x) - metricscore(y))$ is greater than zero". [BELIEF(B), LIME(L), and BayLIME(Ba); D:M denotes Dataset:Model; O refers to Oxford-IIIT Pets dataset. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	D:M	W	p-value	CLES
AOPC Insertion				
B, L	O:I	1229	1.7e-11	0.892
B,Ba	O:I	1188	1.4e-09	0.886
B,L	O:R	1040	2.6e-05	0.756
B,Ba	O:R	1057	1.1e-05	0.753
AOPC Deletion				
B,L	O:I	1231	1.3e-11	0.889
B,Ba	O:I	1184	2.0e-09	0.879
B,L	O:R	1040	2.6e-05	0.758
B,Ba	O:R	1054	1.3e-05	0.753
AUC Insertion				
B,L	O:I	1230	1.5e-11	0.898
B,Ba	O:I	1190	1.2e-09	0.885
B,L	O:R	1051	1.5e-05	0.767
B,Ba	O:R	1076	4.0e-06	0.766
AUC Deletion				
L,B	O:I	1230	1.5e-11	0.894
Ba,B	O:I	1186	1.7e-09	0.883
L,B	O:R	1056	1.2e-05	0.767
Ba,B	O:R	1068	6.1e-06	0.764

7.2 DELETION AND INSERTION GAME

We additionally analyze Insertion and Deletion AUC for fidelity evaluation [Petsiuk et al., 2018] for BELIEF, LIME, and BayLIME. A higher area under the curve (AUC) of the insertion graph indicates higher fidelity of explanations. Conversely, in the deletion procedure, a lower AUC of the deletion graph indicates higher fidelity.

We present the ECDF plots of the AUC insertion and AUC deletion scores for Oxford-IIIT Pets datasets for both models in Figure 6. It can be observed on the top row of Figure 6 that the AUCs for the deletion procedure of BELIEF explanations were lower than those of LIME and BayLIME explanations (Refer Figure S8a in supplementary for ECDF plots for both datasets). We performed the Wilcoxon signed rank tests on the AUCs obtained for all three methods on Oxford-IIIT Pets dataset to confirm this observation (refer Table 2 for test details). Extremely low p-values and high test statistics in all scenarios indicate robust statistical evidence to reject the Null Hypothesis. This confirms that the AUC deletion scores of LIME and BayLIME were much higher than that of BELIEF. Further, the large effect size, i.e., 0.76 to 0.89 proves the practical implications of the same.

Similarly, the higher AUC insertion scores of BELIEF can be seen in the lower row of the ECDF plot in Figure 6 (Refer Figure S8b in supplementary for ECDF plots for both datasets) and the details of statistical test in Table 2. The results from our tests provide robust statistical evidence confirming that the explanations of BELIEF are significantly superior than those of LIME and BayLIME in terms of fidelity and at par with SLICE. The detailed results on both datasets can be found in supplementary material (Table S6).

8 COMPARISON OF BELIEF AND SLICE

8.1 CONSISTENCY COMPARISON

BELIEF and SLICE have almost the same distribution of CCM scores Oxford-IIIT Pets dataset and both model as shown in Figure 3 (refer Figure S3 in supplementary material for ECDF plots of both datasets). However, to confirm that there is no significant difference in their CCM scores, we conducted a Wilcoxon Signed Rank test as shown in Table 3 for Oxford IIIT Pets datatset. We fail to reject the Null Hypothesis ("The median of the differences ($CCM\ score$(BELIEF) $-$ $CCM\ score$(SLICE)) is equal to zero.") as the p-values are much larger than the commonly accepted threshold of 0.05 indicating insufficient statistical evidence to prove that the CCM scores of BELIEF and SLICE are different.

8.2 FIDELITY COMPARISON

We further see that the distribution of the fidelity scores are similar for BELIEF and SLICE as shown in Figure 5 and Figure 6 for Oxford-IIIT Pets dataset for both Inception V3 and ResNet50 models (refer Figure S7 and Figure S8 in supplementary for ECDF plots for both datasets). The high p-values observed in Table 4 which are much greater than the commonly accepted threshold of 0.05 indicate that we fail to reject the Null Hypothesis. Hence, we conclude that there is not enough statistical evidence to prove that the fidelity scores of BELIEF and SLICE are different. (Refer Table S7 in supplementary for test details on both datasets).

Table 3: Wilcoxon Signed Rank test results comparing CCM scores of BELIEF(B) and SLICE(S) with a two-sided alternative hypothesis. In each test, the null hypothesis H_0 was "The median of the differences ($CCM\ score$(BELIEF) $-$ $CCM\ score$(SLICE)) is equal to zero." and the alternative hypothesis was H_a was "The median of the differences ($CCM\ score$(BELIEF) $-$ $CCM\ score$(SLICE)) is not equal to zero". D:M denotes Dataset:Model, where O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W represents the Test Statistic, and CLES denotes the Common Language Effect Size.

Test	D:M	W	p-value	CLES
B, S	O:I	501	0.19	0.392
B, S	O:R	422	0.37	0.432
B, S	P:I	552	0.42	0.392
B, S	P:R	493	0.17	0.428

8.3 RUNTIME COMPARISON

The main difference between BELIEF and SLICE is that BELIEF uses our proposed novel Sign Entropy regularization. In contrast, SLICE uses the frequentist Ridge Regression with bootstrapping to eliminate features with high Sign Entropy making it slow. Further, the main component that takes the highest time is running the predict function on the perturbed sample images generated around the IE. A larger sample size would require more calls to predict, thus increasing the overall execution time.

We therefore analyze the computation advantage of BELIEF as compared to SLICE. To demonstrate the computational advantage of BELIEF, we ran SLICE for 100 random images from Oxford-IIT Pets and PASCAL VOC datasets. We noted the number of calls to the predict function and the execution time for each image and calculated their Pearson correlation. The Pearson correlation for SLICE using ResNet50 was 0.9959, and for Inception V3, it was 0.9953, proving that our assumption regarding the direct impact of sample size on execution time is valid.

Further, we fixed the sample size for BELIEF at 500 for all our experiments and were able to achieve comparable results in consistency and fidelity as compared to SLICE (which used ≈ 2500 samples for ResNet50 and ≈ 3000 samples for Inception V3, as shown in Section 8.1 and Section 8.2). We present the sample sizes used by SLICE for all the images for ResNet50 and Inception V3 models in Figure 7. The median number of samples required for SLICE to stabilize the explanations for ResNet50 was 2500 and for Inception

Table 4: Wilcoxon signed rank test results for comparison of BELIEF(B) and SLICE(S). metric(B,S) indicates the test where the null hypothesis H_0 was "The median of the differences ($metricscore$(BELIEF) $-$ $metricscore$(SLICE)) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($metric\ score$(BELIEF) $-$ $metric\ score$(SLICE)) is not equal to zero". AOPC and AUC are the metrics, D:M denotes Dataset:Model; O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	D:M	W	p-value	CLES
AOPC Insertion				
B,S	O:I	590	.65	0.538
B,S	O:R	557	.44	0.535
AOPC Deletion				
B,S	O:I	589	.65	0.537
B,S	O:R	567	.50	0.528
AUC Insertion				
B,S	O:I	589	.65	0.535
B,S	O:R	546	.38	0.546
AUC Deletion				
B,S	O:I	591	.66	0.462
B,S	O:R	553	.42	0.460

V3 was 3000, which are much larger ($\approx 5X$ times) than that of BELIEF. Based on the distribution information, we employed Kernel Density Estimation (KDE) to estimate the probability of SLICE to have a sample size of 500 or less. We used Scott's method [Scott, 2015] of calculating the bandwidth for the same. The probabilities for SLICE to have less than or equal to 500 sample sizes was $5.058e-03$ for ResNet50 and $1.240e-06$ for Inception V3. BELIEF was therefore able to stabilize LIME explanations with a much smaller sample size and lower average execution time as shown in Table 5. While the running time for BELIEF is comparable to LIME and BayLIME, it provides a high consistency comparable to SLICE.

Table 5: Median running time (lower is better) and CCM scores (higher is better) of BELIEF, SLICE, BayLIME and LIME (in seconds per image) for Inception V3 and Resnet50 models. The values are calculated by running the four methods on 100 randomly sampled images from Oxford-IIIT Pets and PASCAL VOC 2007 datasets for both Resnet50 and Inception V3 models. The median Runtime and CCM scores were computed by aggregating values from both datasets.

Method	Runtime ↓	CCM ↑	Runtime ↓	CCM ↑
	Inception V3		ResNet50	
LIME	5.06	0.232	3.43	0.363
BayLIME	5.04	0.312	3.38	0.501
SLICE	50.53	0.999	30.32	0.999
BELIEF	5.04	0.998	3.39	0.999

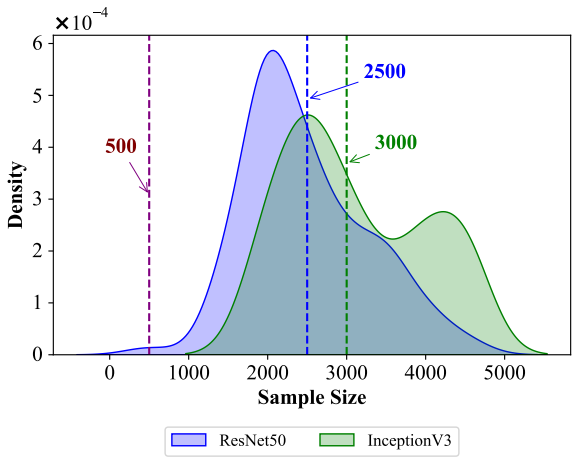


Figure 7: Distribution of sample sizes of SLICE for 100 random images with ResNet50 and Inception V3 models with dotted lines of corresponding colors denoting the respective median values. The sample size of BELIEF is denoted using the blue dotted line at 500. BELIEF uses a much smaller sample size as compared to SLICE to achieve comparable Consistency and Fidelity of Explanations.

9 LIMITATIONS

Fidelity metrics were criticized by Tomsett et al. [2020] highlighting their inconsistency. While we have used the well-known metrics (as discussed in Section 5.1), the reliability and consistency of fidelity metrics were beyond the scope of our paper and should be investigated in future works.

10 CONCLUSION AND FUTURE WORK

The proposed approach of Sign Entropy regularization to enforce sparsity of coefficients achieved robustness in sign flips while maintaining the model's predictive power. The application of our proposed regularization method, BELIEF, is also shown for XAI where the approach adeptly estimates and discards superpixels with high sign variability for consistent explanation. Further, BELIEF works without the additional step of feature selection as compared to previous work leading to a considerable gain ($\approx$ 10X) in execution time while maintaining the same level of consistency and fidelity as compared to previous state-of-the-art method. Our results are also supported by statistical tests that provide statistical significance for our claims. The proposed Sign Entropy-based regularization is thus applicable to tabular and image data, proving its versatility for general-purpose regression tasks and explainability. While we demonstrated the effectiveness of BELIEF on tabular and image data, future work can explore its applicability to other use cases where stability of coefficients is crucial, such as finance, healthcare, and Natural Language Processing (NLP).

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Revoti Prasad Bora, Philipp Terhörst, Raymond Veldhuis, Raghavendra Ramachandra, and Kiran Raja. SLICE: Stabilized lime for consistent explanations for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10988–10996, June 2024.

Luis M Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, 140:81–97, 2017.

Melissa Carroll and Linjie Luo. Lecture 10: Bayesian linear regression. `https://www.cs.princeton.edu/courses/archive/spr09/cos513/scribe/lecture10.pdf`, 2009. COS 513: Foundations of Probabilistic Models, Princeton University, Spring 2009.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

Changyao Chen. Rank-biased overlap (rbo). `https://pypi.org/project/rbo/`, 2023. Accessed: 2023-08-08.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

Alicja Gosiewska and Przemyslaw Biecek. Do not trust additive explanations. *arXiv preprint arXiv:1903.11420*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

JL Hodges Jr. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Kaggle. House prices: Advanced regression techniques, 2024. URL `https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data`. Accessed: April 6, 2024.

Michael H Kutner, Christopher J Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 2005.

Gichan Lee and Scott Uk-Jin Lee. Towards reliable software analytics: Systematic integration of explanations from different model-agnostic techniques. *IEEE Software*, 2023.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xiao Zhang, Ji Liu, Haiyan Jiang, Zeyu Chen, and Dejing Dou. G-lime: Statistical learning for local interpretations of deep neural networks using global priors. *Artificial Intelligence*, 314: 103823, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Kenneth O McGraw and Seok P Wong. A common language effect size statistic. *Psychological bulletin*, 111(2):361, 1992.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

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. *Scikit-learn: Machine Learning in Python*, 2011. URL `https://scikit-learn.org/stable/`. Accessed: 2024-02-08.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Ruslan Salakhutdinov. Sta4273: Bayesian decision theory, lecture 2 notes, 2024. URL `https://www.utstat.toronto.edu/~rsalakhu/sta4273/notes/Lecture2.pdf`. Accessed: 2024-11-14.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Sharath M Shankaranarayana and Davor Runje. 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 20*, pages 454–463. Springer, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6021–6029, 2020.

András Vargha and Harold D Delaney. A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132, 2000.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991*, 2019.

Zhao. Baylime: A bayesian local interpretable model-agnostic explanation approach. `https://github.com/x-y-zhao/BayLime`, 2023.

Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in artificial intelligence*, pages 887–896. PMLR, 2021.

Zhengze Zhou, Giles Hooker, and Fei Wang. S-lime: Stabilized-lime for model explanation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2429–2438, 2021.

BELIEF - Bayesian Sign Entropy Regularization for LIME Framework (Supplementary Material)

Revoti Prasad Bora[1] **Philipp Terhörst**[2] **Raymond Veldhuis**[1] **Raghavendra Ramachandra**[1] **Kiran Raja**[1]

[1]Norwegian University of Science and Technology, Gjøvik, Norway
[2]Paderborn University, Paderborn, Germany

A DEFINITIONS

A.1 OVERVIEW OF LIME

LIME is a popular post-hoc model agnostic method for interpreting the predictions of complex machine learning models [Ribeiro et al., 2016]. LIME approximates a complex model locally with a simpler, transparent model (like linear regression or decision trees) called a surrogate model. This surrogate model, since it is transparent, is used to explain individual predictions in the locality. Mathematically, LIME solves the optimization problem as below:

$$\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where, $f(x)$ is the prediction of the complex model, for instance, x, $g(x')$ is the prediction of the surrogate model for a representation x' of instance x, $\pi_x(z)$ is a proximity measure between instance x and z and $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithfully g approximates f in the vicinity of x, weighted by the proximity measure $\pi_x(z)$ and $\Omega(g)$ is a measure of the complexity of the surrogate model.

LIME's optimization aims to find a surrogate model g that approximates the complex model f in the neighbourhood of x and is transparent in nature. The most important aspects in LIME are the choice of the representation x' and the measure of locality $\pi_x(z)$. The authors use a binary vector x' indicating the presence or absence of interpretable components (like words in text or superpixels/segments in images). Further, a weight function is used to give higher weight to instances that are closer to x. This weight function uses an exponential kernel, i.e. $\pi_x(z) = \exp(-Dist(x, z)^2/\sigma^2)$, where $Dist(x, z)$ is the cosine distance between x and z, and σ is a kernel width parameter.

In the context of explaining images, this involves transforming the problem from image to a tabular format. The process has the following main steps, viz. (1) Divide the image into superpixels or segments using segmentation, (2) Generate random perturbation vectors with length equal to a number of superpixels, (3) Perturbing the superpixels and noting the predictions (output probability) (4) Building a surrogate model with perturbation vectors as X and predictions from step 3 as y, and (5) extracting explanations from the surrogate model. This transformation of the problem statement from images into a tabular format is the vital part of how LIME generalizes the extraction of explanation to image classification models.

A.2 EVALUATION METRICS

We use the same consistency metrics as mentioned in [Bora et al., 2024]. Bora et al. [2024] propose two consistency evaluation metrics to address the two aspects of consistency i.e., coefficients' sign flips and the variance in importance ranks of the coefficients of the surrogate model. These metrics are defined as below:

1. **Average Sign Flip Entropy (ASFE)**: This metric measures the variability in the sign of a superpixel across multiple runs. A lower value of ASFE indicates that the concerned superpixel has lower probability of sign flips across multiple

runs. ASFE for model 'M' and explanation technique 'xp' is calculated as below:

$$ASFE_M^{xp} = \frac{1}{n}\sum_{i=1}^{n} H(\text{sign}_i) \tag{2}$$

where,

$$H(\text{sign}_i) = -p_i^+ \log_2(p_i^+) - p_i^- \log_2({}_ip^-)$$

where, $H(\text{sign}_i)$ is the sign entropy of the i^{th} superpixel. The the probabilities of the i^{th} superpixel to be positive or negative are denoted by the terms p_i^+ and p_i^- respectively. Kernel Density Estimation (KDE), owing to its non-parametric nature, is used to estimate p_i^+ and p_i^- for each of the 'n' superpixels. For bandwidth selection Scott's method is used [Scott, 2015]. $ASFE_{Model}^{xp}$ can range between [0,1] such that 0 represents no sign flips and 1 denotes 50% probability of the coefficient to be positive i.e. high sign flips.

2. **Average Rank Similarity (ARS)**: This measure quantifies the consistency in the importance ranks of superpixels across multiple runs. A higher ARS score indicates that the importance ranks of the superpixels have more agreement across multiple runs than a lower ARS score. Rank Biased Overlap (RBO) score [Webber et al., 2010] is used to calculate the ARS across different runs for model 'M' and explanation technique 'xp' as per the equation below.

$$ARS_M^{xp} = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} rbo\_ext(\boldsymbol{R_i}, \boldsymbol{R_j})}{\binom{m}{2}} \tag{3}$$

Here, $\boldsymbol{R_i}$ and $\boldsymbol{R_j}$ denote the ranked coefficient vectors obtained from the i-th and j-th runs, respectively. The function $rbo_ext(\boldsymbol{R_i}, \boldsymbol{R_j})$ calculates the extrapolated Rank-Biased Overlap (RBO) score between these ranked vectors. To compute the RBO scores, we utilized the Python package 'rbo' [Chen, 2023], configuring the persistence parameter (p) to 0.2 to assign greater emphasis to the highest-ranked elements. The denominator term $\binom{m}{2}$ represents the total number of distinct rank list pairs, ensuring that the rank similarities are averaged across all rank list pairs. The metric ARS_{Model}^{xp} varies between 0 and 1, where a value of 1 signifies a perfect agreement in superpixel rankings across runs, whereas 0 indicates a complete lack of correspondence.

3. **Combined Consistency Metric (CCM)**: Sign entropy and variance the importance ranks of superpixels are quatified by the metrics ASFE and ARS respectively. Bora et. al., thus combined both into a consolidated metric to understand and evaluate an XAI system. The combined metric, CCM, is defined as:

$$CCM_M^{xp} = (1 - ASFE_M^{xp}) * ARS_{Model}^{xp} \tag{4}$$

CCM_M^{xp} ranges between [0,1] where 0 denotes low consistency and 1 denotes full consistency in both sign entropy and superpixel importance ranks.

A.3 ADAPTED AREA OVER PERTURBATION CURVE

We employ the adapted Area Over Perturbation Curve (AOPC), introduced in Bora et al. [2024], to assess the fidelity of explanations generated by LIME, BayLIME, SLICE, and BELIEF. Originally proposed by Samek et al. [2016] as an enhancement of the method by Bach et al. [2015], AOPC quantifies the reduction in predicted probability ($\hat{Y}$) as an image undergoes perturbation, where in our case, perturbations are applied to superpixels based on their ranked importance. While AOPC was initially formulated for deletion, it has since been adapted for insertion as well. The modified AOPC metric is formally defined below.

$$AOPC_d = \frac{1}{L+1}\left\langle \sum_{k=1}^{L} \Delta f(x,k) \right\rangle_{p(x)} \tag{5}$$

where, the term $\Delta f(x,k)$ represents the variation in the classifier's output probability after k perturbation steps, either as an increase or a decrease. For deletion of positive superpixels or insertion of negative superpixels, it is computed as $f(x^{(0)}) - f(x^{(k)})$, where $x^{(0)}$ denotes the original, unperturbed image. Conversely, for insertion of positive superpixels or deletion of negative superpixels, $\Delta f(x,k)$ is defined as $f(x^{(k)}) - f(x^{(0)})$, where $x^{(0)}$ corresponds to the fully perturbed (i.e., blurred) image. The level of blurring is determined using the Adaptive-blur technique from Bora et al. [2024].

Further, $x^{(k)}$ refers to the image at step k during the insertion process, whereas in the deletion process, it represents the progressively restored image after k superpixels from the original image have been reintroduced into the blurred background.

The total number of perturbation steps is denoted by L. The notation $\langle . \rangle_{p(x)}$ indicates the expectation over all dataset images, enabling the computation of the average AOPC score across a deep learning model's predictions.

Additionally, d represents the pixel removal strategy, which can follow either the Most Relevant First (MoRF) or the Least Relevant First (LeRF) order. Since our evaluation involves all superpixels, the insertion and deletion procedures yield identical results. Thus, we conducted all experiments using the MoRF strategy and refer to the computed metric simply as AOPC. As AOPC measures the difference in predicted probabilities between the initial and modified images, a higher AOPC score for both insertion and deletion indicates stronger fidelity. This contrasts with traditional insertion and deletion metrics, where higher insertion AUC and lower deletion AUC signify greater fidelity.

B MAP OBJECTIVE WITH ITERATIVE SIGN ENTROPY PRIOR

Our method introduces a prior over coefficients, one that is not based on magnitude alone, but instead constructed using both the mean and variance of each coefficient's posterior distribution. This prior reflects a more Bayesian treatment by taking into account the full distributional behavior (both mean and variance) of the coefficients.

Carroll et.al. [Carroll and Luo, 2009], described the MAP objective for Bayesian Ridge Regression $\hat{\beta}_{\text{MAP}}$ assuming a Gaussian likelihood with homoscedastic noise of $y_n \mid x_n, \beta \sim \mathcal{N}(x_n^\top \beta, \sigma^2)$ and a zero-mean Gaussian prior $\beta_j \sim \mathcal{N}(0, \tau^2)$ over each of d coefficients as below:

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^\top \beta)^2 + \frac{1}{2\tau^2} \sum_{j=1}^{d} \beta_j^2 \right]$$

Letting $\lambda_1 = \frac{1}{2\tau^2}$, we rewrite this as:

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^\top \beta)^2 + \lambda_1 \sum_{j=1}^{d} \beta_j^2 \right]$$

Sign Entropy Prior (Our Contribution): Unlike traditional priors which penalize β_j based on magnitude alone (i.e., mean), we propose to penalize β_j by using both its mean and variance. We augment the model with the proposed prior that iteratively and adaptively penalizes coefficients based on the their Sign Entropy which is computed using the coefficient's posterior distribution. The proposed iterative prior update resembles the Empirical Bayesian methods where hyper-parameters are refined using the posterior information of the previous iteration [Tipping, 2001].

After iteration $t-1$, the posterior of each coefficient can be approximated as:

$$\beta_j \sim \mathcal{N}(\mu_j^{(t-1)}, (\sigma_j^2)^{(t-1)})$$

We define the Sign Entropy as:

$$\mathcal{H}(\mu_j, \sigma_j) = -p_j \log p_j - (1-p_j) \log(1-p_j), \quad \text{where } p_j = \Phi\left(0; \mu_j, \sigma_j\right)$$

The Sign Entropy prior at t^{th} iteration is given as:

$$\pi(\beta_j^{(t)}) \propto \exp\left(-\lambda_2 \cdot \mathcal{H}(\mu_j^{(t-1)}, \sigma_j^{(t-1)})\right) \tag{6}$$

Although the Sign Entropy prior in Equation 6 is defined as a proportional relationship, due to the bounded nature of $\mathcal{H}(\mu_j, \sigma_j)$ between [0,1] (with log base 2), it can be normalized over a finite domain of β_j.

We solve the standard MAP objective on a reduced set of features, where A^t, the active set at iteration t is defined by a threshold on sign entropy:

$$\mathcal{A}^{(t)} = \left\{ j \in \{1, \ldots, d\} \,\middle|\, \mathcal{H}(\mu_j^{(t-1)}, \sigma_j^{(t-1)}) \leq \zeta \right\}$$

We then solve the MAP problem over this active set $\mathcal{A}^{(t)}$ as below:

$$\hat{\beta}^{(t)} = \arg \min_{\beta_j = 0 \text{ for } j \notin \mathcal{A}^{(t)}} \left[\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^\top \beta)^2 + \lambda_1 \sum_{j \in \mathcal{A}^{(t)}} \beta_j^2 \right] \tag{7}$$

This Sign Entropy prior penalizes coefficients whose sign is inconsistent, and acts as a feedback-based prior, reducing sign entropy of coefficients in subsequent iterative updates. The resulting MAP objective is dynamic and evolves during optimization, leading to improved stability in the sign of the coefficients. While traditional priors like Ridge or Lasso penalize based on the coefficient value alone (and not variance), our Sign Entropy prior incorporates posterior uncertainty (i.e., mean and variance of the coefficients) of previous iteration by penalizing sign inconsistency.

C ADDITIONAL PLOTS

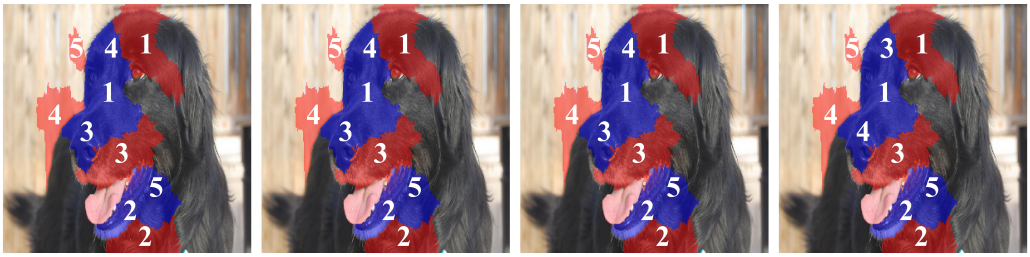


Figure S1: Figure showing the top five positive and negative superpixels of explanations using BELIEF (proposed method) for a random image of the Oxford-IIIT Pets dataset with Inception V3 model for four different runs. The predicted class was Newfoundland, and the prediction probability was 0.46. Blue and red colors denote positive and negative superpixels, and the numbers inside the superpixels specify their importance and rank. There is no inconsistency of superpixels sign i.e., a superpixel deemed as positive in one run is not marked as negative in another and vice-versa. Further, the superpixel importance ranks for both positive and negative superpixels remain stable across all runs.

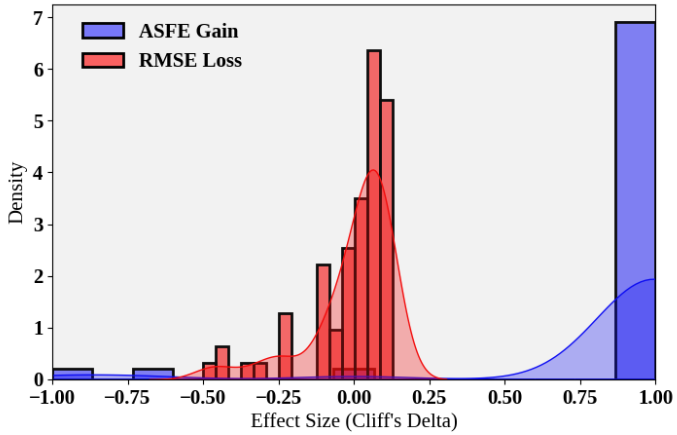


Figure S2: Distribution of effect sizes for Cliff's Delta of ASFE gain and RMSE Loss for the proposed Sign Entropy regularization compared to other well-known approaches for both Energy and Housing datasets. ASFE gain is the decrease in ASFE score (i.e. improvement in Sign Entropy of the coefficients) and RMSE loss is the increase in RMSE score (i.e. the increase in the RMSE of the Linear Regression model). The effect size for ASFE gain is almost always positive and high ('1') except for two cases. The effect size of RMSE loss is either very low or negative. This indicates that our proposed regularization can reduce the sign entropy significantly while keeping the RMSE comparable. We do additional statistical tests to confirm our claims. Please refer Table S9 for details of the conducted statistical tests.

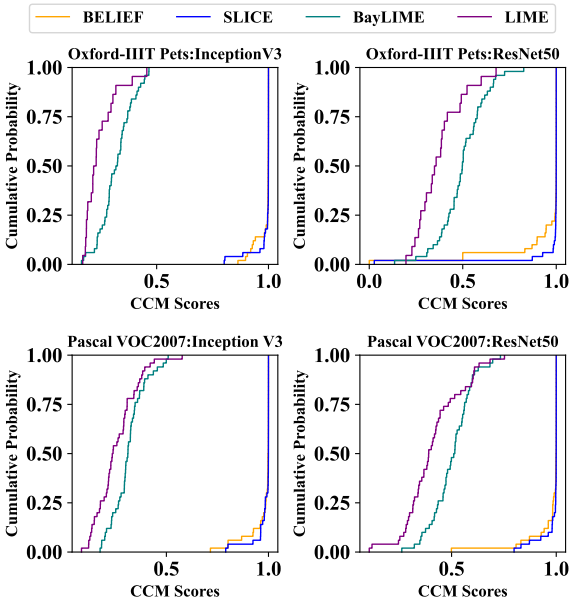


Figure S3: ECDF plot of CCM Scores for BELIEF, LIME, BayLIME and SLICE (higher score is better)

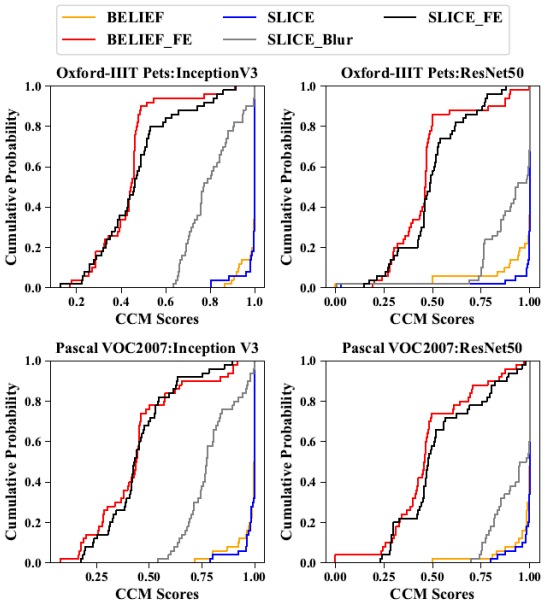


Figure S4: ECDF plot of CCM Scores for BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE and SLICE (higher is better)

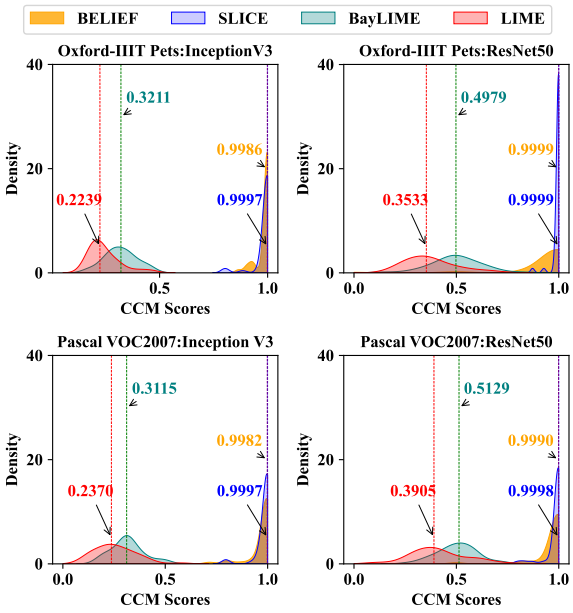


Figure S5: Distribution of CCM Scores for BELIEF, LIME, BayLIME, and SLICE (higher is better).

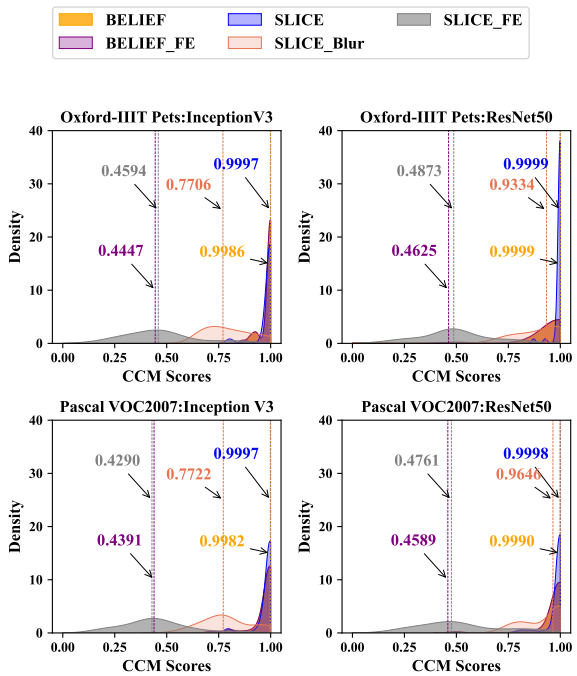
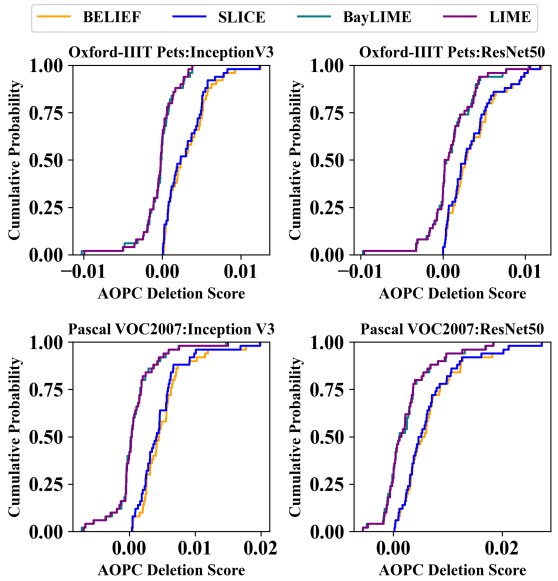
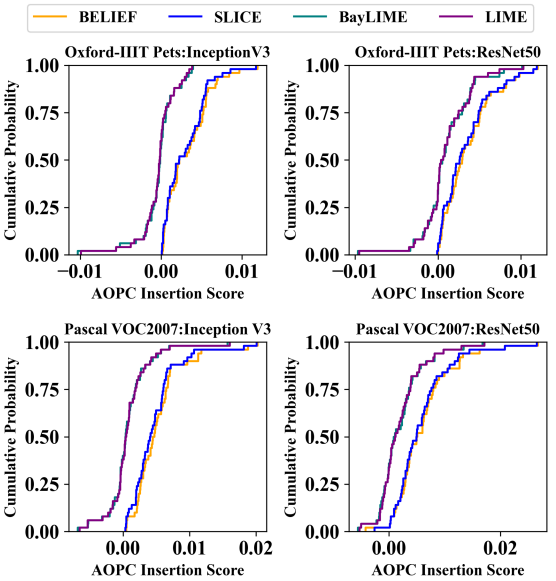


Figure S6: Distribution of CCM Scores for BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE, and SLICE (higher is better).

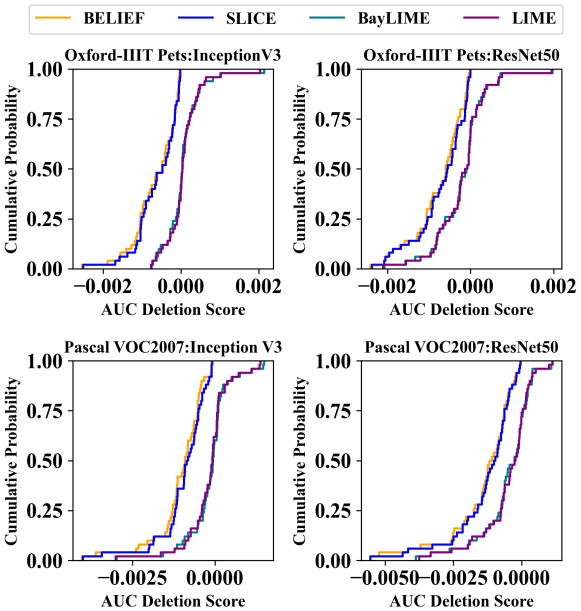


(a) ECDF plots of AOPC deletion scores

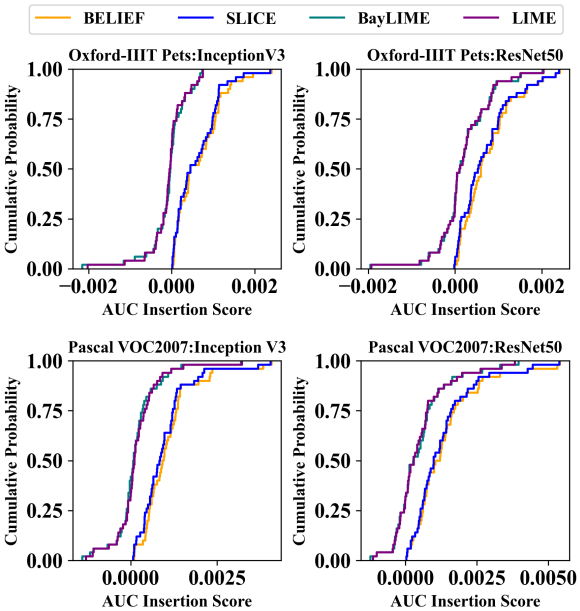


(b) ECDF plots of AOPC insertion scores

Figure S7: ECDF plots of AOPC (Higher AOPC indicates higher fidelity)



(a) ECDF plots of Deletion AUC.(Lower is better)



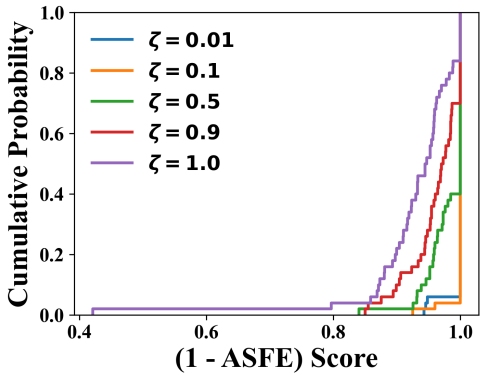
(b) ECDF plots of Insertion AUC.(Higher is better)

Figure S8: ECDF plots of Deletion and Insertion AUC

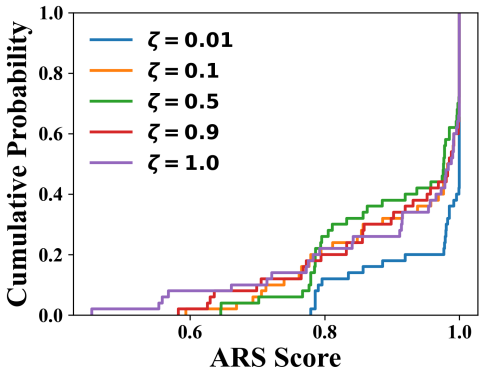
D SENSITIVITY ANALYSIS OF HYPER-PARAMETER ζ

Table S1: Mean CCM scores for different values of ζ on the Oxford-IIIT Pets dataset.

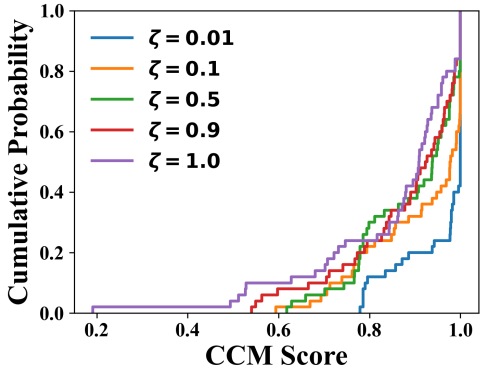
ζ Value	Mean CCM Score
0.01	0.958
0.1	0.915
0.5	0.893
0.9	0.880
1.0	0.851



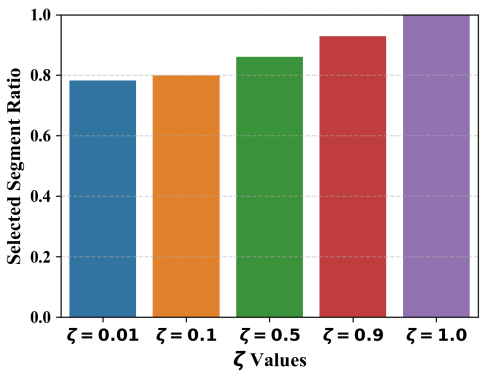
(a) ECDF plot of (1 – ASFE) scores across different ζ variants for the **ResNet-50** model evaluated on 50 randomly selected images from the **Oxford-IIIT Pets** dataset.



(b) ECDF plot of **ARS** scores across different ζ variants for the ResNet-50 model on the same 50-image sample of the Oxford-IIIT Pets dataset.



(c) ECDF plot of **CCM** scores across different ζ variants for the ResNet-50 model on 50 Oxford-IIIT Pets images.



(d) Bar plot showing the **average (mean) ratio of selected segments** obtained with different ζ variants for the ResNet-50 model on the same image set.

Figure S9: Sensitivity analysis for hyper-parameter $\boldsymbol{\zeta}$ of BELIEF on ResNet-50 with 50 Oxford-IIIT Pets images.

In this section we show the results of sensitivity analysis of the hyper-parameter ζ, for the Oxford-IIIT Pets Dataset images on the ResNet50 model. In the plots Figure S9a and Figure S9b, the quantities (1–ASFE) and ARSC decrease as ζ increases. This leads to an overall decrease in CCM scores with increasing ζ, as shown in Table S1 and the ECDF plot of Figure S9c.

Thus, the approach becomes more conservative in selecting features (segments in this case) with a propensity for sign flips as the value of ζ goes down. This is illustrated in Figure S9d, where lowering the value of ζ leads to a decrease in the average (mean) ratio of selected segments.

Therefore, the hyper-parameter ζ should be tuned to balance the trade-off between explainability and feature retention based on the end user's goals. We recommend that in applications where explainability is crucial, the value of ζ be set low based on the acceptable percentage of sign flips; in other situations, it can be relaxed.

E DETAILS OF STATISTICAL TESTS

We performed the Wilcoxon Signed Rank test to ascertain the statistical significance of our results. Additionally, we report the Common Language Effect Size (CLES), which quantifies the proportion of pairs where a value from the first distribution is greater than a value from the second distribution, with an adjustment for tied values McGraw and Wong [1992], Vargha and Delaney [2000].

Table S2: Wilcoxon Signed Rank test results for comparison of CCM scores of BELIEF, BayLIME, and LIME. Here x,y in the test column indicates the test details with x and y. Where x and y are one of B, Ba, and L denotes BELIEF, BayLIME, and LIME respectively. The null hypothesis H_0 was "The median of the differences ($CCM(x) - CCM(y)$) is equal to zero," and the alternative hypothesis H_a was "The median of the differences ($CCM(x) - CCM(y)$) is greater than zero". D:M denotes Dataset:Model where O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	**D:M**	**W**	**p-value**	**CLES**
B, L	O:I	1275	8.9e-16	1.000
B, Ba	O:I	1275	8.9e-16	1.000
B, L	O:R	1267	2.2e-14	0.973
B, Ba	O:R	1227	2.2e-11	0.961
B, L	P:I	1275	8.9e-16	1.000
B, Ba	P:I	1275	8.9e-16	1.000
B, L	P:R	1275	8.9e-16	0.996
B, Ba	P:R	1274	1.8e-15	0.989

Table S3: Wilcoxon Signed Rank test results for comparison of BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE, LIME, and BayLIME for ablation study. Here x,y in the test column indicates the test details with x and y. Where x and y are one of B, Bf, Sb, Sf, L, and Ba denotes BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE, LIME, and BayLIME respectively. The null hypothesis H_0 was "The median of the differences ($CCM(x) - CCM(y)$) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($CCM(x) - CCM(y)$) is greater than zero". D:M denotes Dataset:Model where O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	**D:M**	**W**	**p-value**	**CLES**
B, Bf	O:I	1275	8.9e-16	0.999
B, Sb	O:I	1260	1.2e-13	0.925
B, Sf	O:I	1275	8.9e-16	0.999
B, L	O:I	1275	8.9e-16	1.000
B, Ba	O:I	1275	8.9e-16	1.000
B, Bf	O:R	1261	9.8e-14	0.966
B, Sb	O:R	974	4.4e-04	0.652
B, Sf	O:R	1257	2.2e-13	0.962
B, L	O:R	1267	2.2e-14	0.973
B, Ba	O:R	1227	2.2e-11	0.961
B, Bf	P:I	1272	4.4e-15	0.992
B, Sb	P:I	1267	2.2e-14	0.918
B, Sf	P:I	1275	8.9e-16	0.996
B, L	P:I	1275	8.9e-16	1.000
B, Ba	P:I	1275	8.9e-16	1.000
B, Bf	P:R	1275	8.9e-16	0.988
B, Sb	P:R	904	4.7e-03	0.615
B, Sf	P:R	1269	1.2e-14	0.982
B, L	P:R	1275	8.9e-16	0.996
B, Ba	P:R	1274	1.8e-15	0.989

Table S4: Wilcoxon Signed Rank test results for comparison of BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE, LIME, and BayLIME for ablation study. Here x,y in the test column indicates the test details with x and y. Where x and y are one of B, Bf, Sb, Sf, L, and B denotes BELIEF, BELIEF_FE, SLICE_blur, SLICE_FE, LIME, and BayLIME respectively. The null hypothesis H_0 was "The median of the differences ($CCM(x) - CCM(y)$) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($CCM(x) - CCM(y)$) is greater than zero". D:M denotes Dataset:Model where O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	**D:M**	**W**	**p-value**	**CLES**
B, Bf	O:I	1275	8.9e-16	0.999
B, Sb	O:I	1260	1.2e-13	0.925
B, Sf	O:I	1275	8.9e-16	0.999
B, L	O:I	1275	8.9e-16	1.000
B, Ba	O:I	1275	8.9e-16	1.000
B, Bf	O:R	1261	9.8e-14	0.966
B, Sb	O:R	974	4.4e-04	0.652
B, Sf	O:R	1257	2.2e-13	0.962
B, L	O:R	1267	2.2e-14	0.973
B, Ba	O:R	1227	2.2e-11	0.961
B, Bf	P:I	1272	4.4e-15	0.992
B, Sb	P:I	1267	2.2e-14	0.918
B, Sf	P:I	1275	8.9e-16	0.996
B, L	P:I	1275	8.9e-16	1.000
B, Ba	P:I	1275	8.9e-16	1.000
B, Bf	P:R	1275	8.9e-16	0.988
B, Sb	P:R	904	4.7e-03	0.615
B, Sf	P:R	1269	1.2e-14	0.982
B, L	P:R	1275	8.9e-16	0.996
B, Ba	P:R	1274	1.8e-15	0.989

Table S5: Wilcoxon signed rank test results for comparison of BELIEF (B), LIME (L), and BayLIME (Ba). AOPC(x,y) indicates the test where the null hypothesis H_0 was "The median of the differences ($AOPCscore(x) - AOPCscore(y)$) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($AOPCscore(x) - AOPCscore(y)$) is greater than zero". [D:M denotes Dataset:Model; O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	**D:M**	**W**	**p-value**	**CLES**
Insertion				
AOPC(B,L)	O:I	1229	1.7e-11	0.892
AOPC(B,L)	O:R	1040	2.6e-05	0.756
AOPC(B,L)	P:I	1187	1.5e-09	0.878
AOPC(B,L)	P:R	1098	1.1e-06	0.771
AOPC(B,Ba)	O:I	1188	1.4e-09	0.886
AOPC(B,Ba)	O:R	1057	1.1e-05	0.753
AOPC(B,Ba)	P:I	1171	6.3e-09	0.880
AOPC(B,Ba)	P:R	1028	4.5e-05	0.768
Deletion				
AOPC(B,L)	O:I	1231	1.3e-11	0.889
AOPC(B,L)	O:R	1040	2.6e-05	0.758
AOPC(B,L)	P:I	1187	1.5e-09	0.874
AOPC(B,L)	P:R	1094	1.4e-06	0.775
AOPC(B,Ba)	O:I	1184	2.0e-09	0.879
AOPC(B,Ba)	O:R	1054	1.3e-05	0.753
AOPC(B,Ba)	P:I	1160	1.5e-08	0.876
AOPC(B,Ba)	P:R	1007	1.2e-04	0.768

Table S6: Wilcoxon signed rank test results comparing Insertion and Deletion AUCs of BELIEF (B) with LIME (L) and BayLIME (Ba) using a greater alternative hypothesis. AUC(x,y) denotes a test with null hypothesis H_0 that the median difference in scores between x and y is zero, against an alternative hypothesis H_a of a positive median difference. [D:M signifies Dataset:Model; O for Oxford-IIIT Pets, P for PASCAL VOC, R for ResNet50, and I for Inception V3. W represents the Test Statistic and CLES the Common Language Effect Size.]

Test	**D:M**	**W**	**p-value**	**CLES**
Insertion				
AUC(B,L)	O:I	1230	1.5e-11	0.898
AUC(B,Ba)	O:I	1190	1.2e-09	0.885
AUC(B,L)	O:R	1051	1.5e-05	0.767
AUC(B,Ba)	O:R	1076	4.0e-06	0.766
AUC(B,L)	P:I	1183	2.2e-09	0.872
AUC(B,Ba)	P:I	1169	7.4e-09	0.877
AUC(B,L)	P:R	1113	4.4e-07	0.773
AUC(B,Ba)	P:R	999	1.6e-04	0.763
Deletion				
AUC(L,B)	O:I	1230	1.5e-11	0.894
AUC(Ba,B)	O:I	1186	1.7e-09	0.883
AUC(L,B)	O:R	1056	1.2e-05	0.767
AUC(Ba,B)	O:R	1068	6.1e-06	0.764
AUC(L,B)	P:I	1184	2.0e-09	0.872
AUC(Ba,B)	P:I	1156	2.1e-08	0.874
AUC(L,B)	P:R	1098	1.1e-06	0.775
AUC(Ba,B)	P:R	997	1.8e-04	0.765

Table S7: Wilcoxon signed rank test results for comparison of BELIEF(B) and SLICE(S). metric(B,S) indicates the test where the null hypothesis H_0 was "The median of the differences ($metricscore$(BELIEF) − $metricscore$(SLICE)) is equal to zero," and the alternative hypothesis was H_a was "The median of the differences ($metric\ score$(BELIEF) − $metric\ score$(SLICE)) is not equal to zero". AOPC and AUC are the metrics, D:M denotes Dataset:Model; O refers to Oxford-IIIT Pets and P refers to PASCAL VOC datasets. R denotes ResNet50 and I denotes Inception V3 models. W denotes the Test Statistic and CLES denotes the Common Language Effect Size.

Test	**D:M**	**W**	**p-value**	**CLES**
AOPC Insertion				
AOPC(B,S)	O:I	590	.65	0.538
AOPC(B,S)	O:R	557	.44	0.535
AOPC(B,S)	P:I	597	.70	0.557
AOPC(B,S)	P:R	559	.45	0.522
AOPC Deletion				
AOPC(B,S)	O:I	589	.65	0.537
AOPC(B,S)	O:R	567	.50	0.528
AOPC(B,S)	P:I	596	.69	0.557
AOPC(B,S)	P:R	548	.39	0.521
AUC Insertion				
AUC(B,S)	O:I	589	.65	0.535
AUC(B,S)	O:R	546	.38	0.546
AUC(B,S)	P:I	597	.70	0.558
AUC(B,S)	P:R	555	.43	0.525
AUC Deletion				
AUC(B,S)	O:I	591	.66	0.462
AUC(B,S)	O:R	553	.42	0.460
AUC(B,S)	P:I	595	.69	0.444
AUC(B,S)	P:R	547	.39	0.478

Table S8: Median ASFE scores and RMSE of our proposed Sign Entropy regularization and other approaches. Lower ASFE and RMSE scores are better. OLS does not have a regularization term and ARD does not have $lambda\_init$ hyper-parameter. Therefore, we conducted the experiments without applying the regularization hyper-parameter settings (0.1, 0.5, and 1), and we denote this scenario using the same values of ASFE and RMSE for R1, R.5, and R1 in OLS and ARD.

M	ASFE ↓						RMSE ↓					
	Proposed	Lasso	Ridge	Bayesian Ridge	ARD	OLS	Proposed	Lasso	Ridge	Bayesian Ridge	ARD	OLS
Housing Price Dataset												
R.1	0.149	0.474	0.46	0.451	0.427	0.474	0.319	0.316	0.311	0.310	0.432	0.317
R.5	0.15	0.465	0.412	0.439	0.427	0.474	0.293	0.294	0.284	0.290	0.432	0.317
R1	0.149	0.462	0.398	0.443	0.427	0.474	0.343	0.345	0.307	0.337	0.432	0.317
Energy Appliances Dataset												
R.1	0.004	0.029	0.183	0.198	0.16	0.192	0.502	0.518	0.472	0.473	0.47	0.469
R.5	0.004	0.000	0.201	0.194	0.16	0.192	0.502	0.595	0.474	0.473	0.47	0.469
R1	0.000	0.000	0.262	0.184	0.16	0.192	0.503	0.596	0.478	0.475	0.47	0.469

Table S9: Kolmogorov-Smirnov (KS) test results comparing proposed Sign Entropy regularization with other methods. For each test, the null hypothesis H_0 was "The distribution of the RMSE score of our proposed regularization is the same as the compared method," and the alternative hypothesis H_a was "The distributions are different." KS Statistic refers to the maximum distance between cumulative distributions, and p-value indicates the probability of observing the result under H_0. Results are grouped by dataset: Energy (E) and Housing (H). All p-values are larger than the commonly accepted threshold of 0.05 except for the proposed method vs. ARD for Housing Dataset (highlighted in red). However, as seen from the RMSE density plots in fig. 2 and table S8, the RMSE of ARD in this case is much higher than other methods. Thus, we conclude that there is no statistically significant **increase** in the RMSE score due to our proposed Sign Entropy regularization.

Test	**KS Statistic**	**p-value**
Energy Dataset (E)		
Proposed vs OLS	0.200	0.731
Proposed vs ARD	0.200	0.731
Proposed vs Bayesian Ridge	0.133	0.825
Proposed vs Lasso	0.200	0.332
Proposed vs Ridge	0.111	0.948
Housing Dataset (H)		
Proposed vs OLS	0.110	0.958
Proposed vs ARD	0.550	0.000
Proposed vs Bayesian Ridge	0.050	0.999
Proposed vs Lasso	0.040	1.000
Proposed vs Ridge	0.050	0.999