2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 979-8-3503-0918-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WI-IAT59888.2023.00036

In-Training Explainability Frameworks: A Method to Make Black-Box Machine Learning Models More Explainable

Cagla Acun
Web Mining and Knowledge Discovery Lab
University of Louisville
Louisville, KY, USA
a0acun01@louisville.edu

Olfa Nasraoui
Web Mining and Knowledge Discovery Lab
University of Louisville
Louisville, KY, USA
olfa.nasraoui@louisville.edu

***Abstract*—Despite ongoing efforts to make black-box machine learning models more explainable, transparent, and trustworthy, there is a growing advocacy for using only inherently interpretable models for high-stake decision making. For instance, post-hoc explanations have recently been criticized because they learn surrogate white-box (explainer) models that, while optimized to approximate the original predictive model, remain different from the latter. Moreover, the post-hoc models necessitate a post-hoc training phase at prediction time, that adds to the computational burden. In this paper, we propose two novel explainability approaches that make black-box models more explainable, which we call pre-hoc explainability and co-hoc explainability. Our goal is to maintain the black-box model's prediction accuracy while benefiting from the explanations that come with an inherently interpretable white-box model, and without the need for a post-hoc training phase at prediction time. In contrast to post-hoc methods, the black-box model training phase is guided by explanations that are used as a regularizer. Our experiments demonstrate the advantages of our proposed technique on three real-life datasets, in terms of fidelity, without compromising accuracy.**

***Index Terms*—Explainability in Artificial Intelligence, XAI**

I. Introduction

Machine learning models are increasingly being used to support decision-making in various fields, from personalized medical diagnosis to credit risk assessment and criminal justice. However, the increasing reliance on powerful black-box models raises concerns about their transparency, interpretability, and trustworthiness [1] [2] [3]. The ability to understand why a model made a particular prediction is crucial to supporting auditing models, detecting potential biases and errors, and, in turn, supporting model accountability and fairness.

Several approaches have been proposed to explain black-box models, ranging from local methods that provide explanations for individual predictions to global methods that aim to capture the model's overall behavior. Post hoc explanations, such as LIME (Local Interpretable Model-Agnostic Explanations) [4], SHAP (Shapley Additive Explanations) [5], and Grad-CAM (Gradient Weighted Class Activation Mapping) [6], have gained popularity in recent years as a way to explain black-box models by perturbing the input data and learning a surrogate model that approximates the original model's behavior locally. Although these methods can be effective in generating explanations, they have been criticized for several reasons. First, the explanations may not reflect the true mechanisms of the original model, but rather a simplified version that is easier to interpret [7]. Second, the surrogate model may not be faithful to the behavior of the original model in some cases, leading to potentially misleading explanations and being open to adversarial attacks [8]. Third, the perturbation of the input data can alter the semantics of the features, rendering the explanations invalid or misleading and unstable explanations that arise with models already trained [9] [10].

To address these limitations, some researchers have proposed the use of inherently interpretable models, such as decision trees or linear models, instead of black-box models for high-stakes decision-making [11]. However, this approach may come at the cost of reduced prediction accuracy, as interpretable models may not be able to capture the complexity of some datasets as well as black-box models. Moreover, the use of interpretable models does not solve the problem of explaining black-box models that are already in use.

In this paper, we propose two novel approaches to enhancing the explainability of black-box models, which we call *pre-hoc explainability* and *co-hoc explainability*. Our approach aims to incorporate explanations derived from an inherently interpretable white-box model into the original model's learning stage without compromising its high prediction accuracy. Unlike post-hoc explanations, our approach does not rely on input perturbation or secondary model learning and thus avoids the potential pitfalls of surrogate modeling. Instead, we leverage the insights provided by a white-box model to guide the training of the black-box model in a way that preserves its accuracy while enhancing its global interpretability. We show that our approach outperforms traditional black-box and white-box models on several benchmark datasets and offers a promising direction for making machine learning models more transparent and trustworthy. Our contributions are summarized below:

- We propose two novel approaches to enhancing the

979-8-3503-0918-8/23/$31.00 ©2023 IEEE
DOI 10.1109/WI-IAT59888.2023.00036

explainability of black-box models, called *pre-hoc explainability* and *co-hoc explainability*, which leverage the insights provided by an inherently interpretable white-box model to guide the training of the black-box model in a way that preserves its accuracy while enhancing its interpretability.
- Unlike post-hoc explanations, our approaches do not rely on input perturbation or post-secondary model learning, and thus avoid the potential pitfalls of surrogate modeling. This makes it more scalable, robust, and reliable in practice.
- We demonstrate the effectiveness of our approaches on several real-world benchmark datasets, showing that it outperforms traditional black-box in terms of fidelity.
- We provide a theoretical analysis of our approaches, showing that it can be seen as a form of regularized learning that balances the trade-off between accuracy and interpretability.

II. Related Work

The majority of the existing work on explainable AI has focused on either developing post hoc explanation methods for black-box models or building models that are explainable by design. Post-hoc techniques analyze trained models to provide explanations for individual predictions [5] [4] [12], either with model-specific methods based on input perturbations or model-agnostic explainer models. However, post hoc approaches have been criticized for potential discrepancies between the explainer and the black-box model [8] [13]. On the other hand, model-specific explainability has its own limitations as it requires individual methods and implementations for each different black-box model.

In contrast, research on enhancing explainability through model training is more limited. Only a few methods have explored using interpretable models to directly guide black-box training for higher explainability. Using tree regularization [14] to train deep time-series models, with the aim of human-simulability [15]. Other works proposed training models with latent explainability, but they still rely on post hoc explanations [16] [17]. An alternative approach is to use a game-theoretic approach between predictor and explainer [18], [19]. By using a cooperative game, they optimize the explainer for locality, specifically for sequential data. [20] used a regularization approach to nudge black-box models toward relying more on interpretable features, but their explanations remain post-hoc, specifically optimized for LIME's neighborhood-based fidelity, which has to be computed *at prediction time*. In fact, their goal is to improve the *quality* of *post-hoc* explanations of the model, thus they do not attempt to solve the same problem as ours, as we do not rely on post-hoc explanations. Another line of work designed to learn the latent concept-based explanations implicitly during training, which eliminates the requirement of post-hoc explanation generation techniques [21]. Because the concepts must be learned using either external annotation or self-supervision, e.g. using auto-encoders from the input features, this approach is limited to special input types like images or domains with available external supervision.

Overall, research on enhancing model explainability highlights the need for further work on optimization *during training*, and model-agnostic methods to improve global explainability. Our approach addresses this need by directly injecting global interpretability into black-box learning, *at training time*, through an interpretable explainer model, that does *not* require additional post-hoc computation at prediction time.

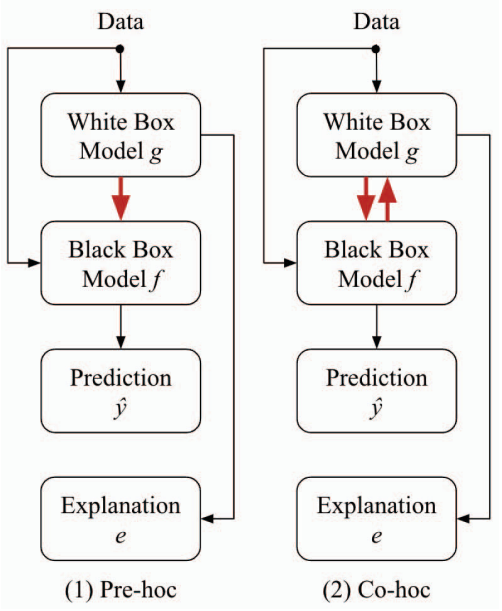


Fig. 1: Proposed Explainability Frameworks

III. Problem Formulation

Let $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\mathcal{N}} \subset \mathcal{Z}$ be a sample from a distribution $\mathcal{D}$ in a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance and $\mathcal{Y}$ is the label set. We learn a differentiable predictive function $f \in \mathcal{F} : \mathcal{X} \to \mathcal{Y}$ together with a transparent function $g \in \mathcal{G} : \mathcal{X} \to \mathcal{Y}$ defined over a functional class $\mathcal{G}$. We refer to functions f and g as the predictor and the explainer, respectively, throughout the paper. $\mathcal{G}$ is strictly constrained to be an inherently explainable functional set, such as a set of linear functions or decision trees. We assume that we have a distance function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that $d(y, \hat{y}) = 0 \longleftrightarrow y = \hat{y}$, which measures the point-wise similarity between two probability distributions in $\mathcal{Y}$ and can be used to optimize f and g.

Our idea is, instead of learning a post hoc white-box model, to learn a model that is explainable from the start and then let this explainer model guide the predictor model. To accomplish this goal, there are several ways. We design two different frameworks; (1) A Pre-Hoc Explainable Predictive Framework, where the white box model regularizes the black box model for optimized fidelity and (2) A Co-hoc Explainable Predictive Framework, where white-box and black-box models are optimized simultaneously with a shared loss function that enforces fidelity. See Figure 1.

IV. Proposed Explainability Frameworks

In this section, we define our fidelity objective function and show two different implementations to enforce fidelity and present them as two novel frameworks, as stated in Section III problem formulation.

We use the explainer function $g \in \mathcal{G}$ to guide the predictor f by means of distance measures globally. We define global interpretability by measuring how close f is to a family $\mathcal{G}$ over N number of batches in point-wise fashion, see Figure 2.

A. Enforcing Fidelity

Definition 1 (Fidelity Objective Function)**.** *Given an inherently interpretable white-box model g with parameters ϕ, let its predictions result in a probability distribution p^{ϕ}. Given the black-box, f with parameters θ, let its predictions result in probability distribution p^{θ} over K classes $y \in \mathcal{Y} = \{1,2,..,K\}$. We propose a fidelity objective function, which measures the point-wise probability distance between p^{ϕ} and p^{θ}, which are respectively the outputs of g and f for all given input data $\mathcal{X}$. Our global distance metric is as follows:*

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} D\left(f\left(x_i\right), g\left(x_i\right)\right), \tag{1}$$

where function D is a divergence distance measurement, Jensen-Shannon divergence [22]. We aim to use D_{JS}, Jensen-Shanon divergence, to measure the point-wise deviation of the predictive distributions f_{θ} and g_{ϕ}.

Denote by $\mathcal{P}$ the set of probability distributions. Kullback-Leibler divergence (KL). $\mathrm{KL} : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]$ is a fundamental distance between probability distributions in $\mathcal{D}$ [23], defined by:

$$D_{\mathrm{KL}}(p||q) := \int p \log \frac{p}{q} \, \mathrm{d}\mu, \tag{2}$$

where p and q denote probability measures P and Q with respect to μ.

Let $p, q \in \Delta^{K-1}$ have the corresponding weights $\pi = [\pi_1, \pi_2]^T \in \Delta$. Then, the Jensen-Shannon divergence between p and q is given by

$$\begin{aligned} D_{\mathrm{JS}}\left(p, q\right) &:= H(m) - \pi_1 H\left(p\right) - \pi_2 H\left(q\right) \\ &= \pi_1 D_{\mathrm{KL}}\left(p \| m\right) + \pi_2 D_{\mathrm{KL}}\left(q \| m\right), \end{aligned} \tag{3}$$

with H the Shannon entropy, and $m = \pi_1 p + \pi_2 q$. Unlike the Kullback-Leibler divergence ($D_{\mathrm{KL}}\left(p \| q\right)$), JS is symmetric, bounded, and does not require absolute continuity.

The fidelity objective function, L_{JSD}, is calculated using the Jensen-Shannon divergence (JS), as follows:

$$\mathcal{L}_{\mathrm{JS}}\left(x_{1:N}, f_{\theta}, g_{\phi}\right) := D_{\mathrm{JS}}(\hat{y}^{\phi}, \hat{y}^{\theta}) \tag{4}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{JS}}\left(x_{1:N}, f_{\theta}, g_{\phi}\right) := \frac{1}{2}(&D_{\mathrm{KL}}(\hat{y}^{\phi} \parallel \frac{(\hat{y}^{\phi} + \hat{y}^{\theta})}{2}) \\ &+ D_{\mathrm{KL}}(\hat{y}^{\theta} \parallel \frac{(\hat{y}^{\phi} + \hat{y}^{\theta})}{2})) \end{aligned} \tag{5}$$

Our goal is to learn the black-box predictive model f_{θ} to optimize fidelity to an inherently explainable g_{ϕ}.

Substituting Eq. 2 into $\mathcal{L}_{JS}$ (Eq. 5), we obtain:

$$\begin{aligned} \mathcal{L}_{JS}(\hat{y}^{\phi} \parallel \hat{y}^{\theta}) = \frac{1}{2}(&\sum_{i=1}^{N} \ln\left(\frac{\hat{y}^{\phi}}{\hat{y}^{\theta}}\right) \hat{y}^{\phi} \\ &+ \sum_{i=1}^{N} \ln\left(\frac{\hat{y}^{\theta}}{\hat{y}^{\phi}}\right) \hat{y}^{\theta}) \end{aligned} \tag{6}$$

Our proposed fidelity objective function has three distinct regularization properties that we explain below.

a) Bounded Regularizer: The Jensen–Shannon divergence distance is always bounded, i.e.,

$$0 \leq \mathrm{JS}(p : q) \leq \log 2, \tag{7}$$

Since the square root of the JS yields a metric distance satisfying the triangular inequality [24]. Thus, lower and upper bounds become

$$0 \leq D_{\mathrm{JS}}(p : q) \leq \sqrt{\log 2}. \tag{8}$$

b) Symmetry Preserving Regularizer: The Jensen Shannon divergence is symmetric w.r.t. two input variables if swapping them does not change the distance. For instance, D_{JS} is symmetric w.r.t. p and q if and only if $D_{\mathrm{JS}}(p;q) = D_{\mathrm{JS}}(q;p)$ for all values of p and q. JS is symmetry preserving if the corresponding weights $\pi = [\pi_1, \pi_2]$ are selected as $\pi = \left[\frac{1}{2}, \frac{1}{2}\right]$.

c) Differentiable Regularizer: Our fidelity loss implements a differentiable regularizer to enforce fidelity between the predictor model and the explainer model, which is used to derive explanations for the predictor model. The regularizer is based on the Jensen-Shannon divergence (JS) between the probability distributions of the explainer model and the predictor model outputs.

Thus, the regularizer is differentiable, which means that it can be easily incorporated into the training process of the predictor model using standard backpropagation techniques. By minimizing the JS between the two distributions, the regularizer encourages the predictor to produce similar probability distributions to the explainer model, thereby ensuring that the explanations derived from the explainer model are more accurate and trustworthy.

B. Pre-hoc Explainability Framework

We formulate the framework in Figure 1 (a) into a modified learning objective to obtain the Pre-hoc explainability as follows

$$\mathcal{L}_{Pre-hoc} = \mathcal{L}_{BCE} + \lambda_1 D_{\mathrm{JS}} + \lambda_2 \mathcal{L}_2, \tag{9}$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy loss, λ_1 is an explainability regularization coefficient that controls the smoothness of the new representation and the trade-off between explainability and accuracy, while λ_2 coefficient for standard $\mathcal{L}_2$ regularization of model parameters θ that aims to avoid overfitting and exploding gradients.

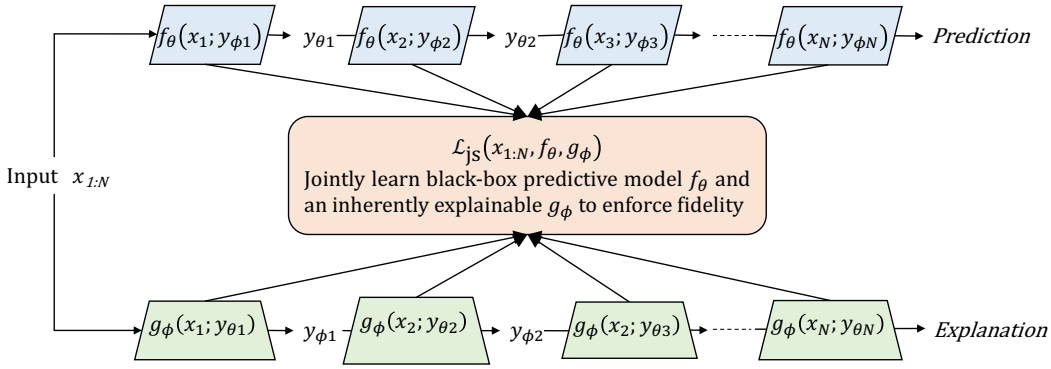


Fig. 2: Training Phase of Co-hoc Explainability Framework

$$\mathcal{L}_{Pre-hoc}(\theta,\phi,X,y,) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} -y_\theta \log(\hat{y}_\theta) + (1-y_\theta)\log(1-\hat{y_\theta})}_{\text{Predictor Accuracy}}$$
$$+\underbrace{\lambda_1 \frac{1}{2}(\sum_{i=1}^{N} \ln\left(\frac{\hat{y}_\phi}{\hat{y}_\theta}\right)\hat{y}_\phi + \sum_{i=1}^{N} \ln\left(\frac{\hat{y_\theta}}{\hat{y}_\phi}\right)\hat{y_\theta})}_{\text{Fidelity}} + \underbrace{\lambda_2 \sum_i \theta_i^2}_{\text{L2 Regularization}} \quad , \tag{10}$$

$\mathcal{L}$ consists of cross-entropy loss and a fidelity regularization term along with a $\mathcal{L}_2$ regularization term.

Since the explanation e is provided by the white box model that is inherently interpretable, the transparency is considered high when the explanatory white box model outputs $\hat{y}_\phi$ are similar to the regularized model outputs $\hat{y}_\theta$. This is captured by D_{JS}, which is term 2, Fidelity, in the proposed objective function, $\mathcal{L}_{\text{Pre-hoc}}$ (eq. 9). While the objective function is to learn the predictions, we give greater importance to the predictions that are similar to the white-box predictions and penalize those that are not similar.

C. Co-hoc Explainability Framework

We formulate this framework (Figure 2) into a modified learning objective to obtain the Co-hoc explainability as follows

Definition 2 (Co-hoc Fidelity Objective Function)**.** *Given an inherently interpretable white-box model g with parameters ϕ, let its predictions result in a probability distribution p^ϕ and given the black-box model f with parameters θ, let its predictions result in probability distribution p^θ over K classes $y \in \mathcal{Y} = \{1,2,..,K\}$. We propose a Co-Learning Explainability Framework, where f_θ and g_ϕ are jointly learned, given p^ϕ and p^θ, respectively, as inputs. We use an added distance function (eq. 1) as a regularization for the objective function to guide the co-learning process. Our global distance metric is the same as the Definition 1, and the combined Co-hoc loss function is given by*

$$\mathcal{L}_{Co-hoc}(\theta,\phi,X,y,) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} -y_n \log(\hat{y}_{\theta n}) + (1-y_n)\log(1-\hat{y_{\theta n}})}_{\textit{Predictor Accuracy}}$$
$$+\underbrace{\frac{1}{N}\sum_{i=1}^{N} -y_n \log(\hat{y}_{\phi n}) + (1-y_n)\log(1-\hat{y_{\phi n}})}_{\textit{Explainer Accuracy}}$$
$$+\underbrace{\lambda_1 \frac{1}{2}(\sum_{i=1}^{N} \ln\left(\frac{\hat{y}_\phi}{\hat{y_{\theta n}}}\right)\hat{y}_{\phi n} + \sum_{i=1}^{N} \ln\left(\frac{\hat{y_{\theta n}}}{\hat{y}_n}\right)\hat{y_{\theta n}})}_{\textit{Fidelity}}$$
$$+ \underbrace{\lambda_2 \sum_i \theta_i^2}_{\textit{Regularization 1}} + \underbrace{\lambda_3 \sum_i \phi_i^2}_{\textit{Regularization 2}} \quad , \tag{11}$$

which contains binary cross-entropy and fidelity regularization terms, along with other regularization terms; Regularization 1 discourages exploding gradients, and Regularization 2 encourages the sparsity of the explainer model.

The primary distinction between Co-hoc and Pre-hoc lies in the *joint* optimization of predictor f_θ and explainer g_ϕ through simultaneous stochastic gradient descent with mini-batches, see Figure 2.

V. Experiments

We conduct experiments that aim to answer the following research questions:

RQ1: Can we maintain an accuracy that is higher than the explainer model even if it is lower than the baseline BB predictor model; while having explanations from g_ϕ (since g_ϕ was used to guide f_θ?)

RQ2: How good is our regularized predictor model f_θ at mimicking the explainer model g_ϕ?

RQ3: How does λ_1 affect the fidelity and accuracy trade-off?

RQ4: What are the differences between the pre-hoc and co-hoc frameworks?

A. Experimental Settings

a) Datasets: We experimented with three publicly accessible real-world datasets. All three datasets used in a binary classification setting. Movielens 100k movie ratings, has 100,000 ratings based on 1000 users on 1700 movies. MovieLens 1M movie ratings, has 1 million ratings based on 6000 users on 4000 movies. For Movielens datasets [25], the classification target is the movie rating. Our goal is to learn like or dislike a movie. The target is discretized into liked and disliked; 1 is the class label for a rating of 3 and above; 0 is the class label for a rating of less than 3. FICO HELOC dataset [26] contains 10,459 anonymized information about home equity line of credit (HELOC) applications made by real homeowners. The target variable is risk performance, which predicts whether the homeowner qualifies for a line of credit or not.

b) Evaluation Protocols: To assess the classification accuracy, we use the Area under the ROC Curve, AUC(f_θ, $\hat{y}$). Each dataset is split randomly into training, validation, and test sets in the ratio 80:10:10. After training on every batches with a learning rate of 0.001, AUC is calculated on the validation and test datasets. We measure all the metrics on a held-out test set. All models are trained with L_2 regularization until validation accuracy is stabilized for at least ten epochs.

Fidelity, also known as descriptive accuracy [27], measures how accurately an explanation method can mimic the behavior of a black-box classifier in terms of assigning class labels to data records. We use AUC(f_θ, g_ϕ) to evaluate the fidelity. Our baseline for fidelity is the AUC of the original black-box predictor model and the explainer model, which can also be considered as a post-hoc explainability score without any optimization.

c) Baselines: We compared our Pre-hoc and Co-hoc predictor models with their original black-box (BB) version. The black-box model is Factorization Machines [28] as it is widely used for classification, regression, and recommendation tasks. The explainer white-box model (WB) is a sparse logistic regression model, which is inherently explainable, and thus provides the explanation.

d) Parameter Settings: We implemented our proposed methods based on PyTorch. All models are learned by optimizing the binary cross entropy and with Adam [29], which is an extension to stochastic gradient descent. Batch size is selected as 64,2056,64 respectively, for ML-100K, ML-1M, and HELOC datasets, which are the optimal batch size for each dataset. We tested λ_1 for $\{0.01, 0.1, 0.25, 0.5, 0.75, 1\}$. The regularization weight of the loss function is estimated using a mini-batch. We pick the best regularization weight for each dataset using the validation set and use that for the final evaluation. The final evaluation is done by retraining the models using their chosen configurations and evaluating them on the test set.

TABLE I: Model comparison in terms of prediction accuracy and fidelity of explainability on the three real-world datasets, two interaction datasets, ML100k and ML1M, and one tabular, HELOC dataset. All evaluation metrics are computed with, respectively, for datasets $\lambda_1 = 0.75, 0.5, 1$ and batch size $n = 64, 2056, 64$. The explainer is the white-box model, BB is the predictor black-box model. The best results are in **bold**. Higher AUC and Fidelity is better.

Dataset	ml-100k		ml-1M		HELOC	
Model	AUC	Fidelity	AUC	Fidelity	AUC	Fidelity
Explainer WB	0.7655	-	0.7882	-	0.7616	-
Original BB	0.7784	0.8287	**0.8078**	0.8875	0.7703	0.7728
Pre-hoc BB	0.7801	0.9094	0.8033	0.9404	0.7698	0.8454
Co-hoc BB	**0.7816**	**0.9194**	0.8036	**0.9484**	**0.7743**	**0.8572**

VI. Results and Discussion

RQ1: For the predictor model f_θ, can we maintain an accuracy that is higher than the explainer model g_ϕ if lower than the original f; while having explanations from g_ϕ (since g_ϕ was used to guide the f_θ)?

The experimental results in Table I, Figure 3, Figure 4 provide insights into this question. Comparing the accuracy scores between the predictor model and the explainer model, we observe that the predictor model for both Pre-hoc predictor and Co-hoc predictor consistently achieves significantly higher accuracy scores than the explainer model g_ϕ (p-value $< .05$), even with the $\lambda_1 = 1$, which is the highest coefficient for optimizing fidelity, for each dataset and the both proposed models.

These results clearly demonstrate that the proposed Pre-hoc and Co-hoc predictors maintain higher accuracy compared to the explainer model g_ϕ baseline while achieving improved fidelity in mimicking the behavior of the explainer model g_ϕ.

RQ2: How good is our regularized predictor model f_θ in mimicking the explainer model g_ϕ?

On the ml-100k dataset, Pre-hoc predictor and Co-hoc predictor achieve fidelity scores of 0.9094 and 0.9194, respectively, 9.7% and 10.9% increase by outperforming the original black-box predictor fidelity score of 0.8287. This indicates that the proposed models better capture the behavior of the explainer model compared to the baseline predictor model. Similarly, on the ml-1M dataset, Pre-hoc predictor and Co-hoc predictor achieve a fidelity score of 0.9404 and 0.9484, outperforming the original fidelity score of 0.8875 by improving 5.9% and 6.8%. Moreover, in the HELOC dataset, improvement in the fidelity score is respectively, 9.3% and 10.9% for Pre-hoc predictor and Co-hoc predictor.

In summary, the experimental findings support the notion that the proposed models, Pre-hoc and Co-hoc predictors, are more successful in mimicking the behavior of the explainer model compared to the baseline black-box model. This demonstrates the effectiveness of the proposed techniques in

enhancing the fidelity of the predictor model while maintaining high accuracy.

Importantly, these improvements in fidelity are achieved without any decrease in the accuracy score. The proposed predictor models maintain a similar level of accuracy compared to the original black-box predictor. The trade-off between fidelity and accuracy will be explored and discussed further in the next research question.

RQ3: How does lambda affect the fidelity and accuracy trade-off?

As we can see in Table II, the regularization hyperparameter λ_1 plays a crucial role in balancing the fidelity and accuracy trade-off in our models. A value of $\lambda_1 = 0$ indicates that there is no regularization of the fidelity, while $\lambda_1 = 1$ signifies an equal weight of the fidelity and accuracy in the objective function. The impact of λ on the results is noteworthy. When λ_1 is set to 0.01 and 0.1, there is no noticeable difference in the Fidelity AUC metric across all datasets. However, as we increase the value of λ_1, we observe a consistent improvement in the fidelity results. In particular, when λ_1 is set to 1, we achieve almost perfect fidelity scores, indicating a high level of agreement between the explainer model g_ϕ and the regularized predictor model f_θ. This demonstrates the effectiveness of the regularization approach in mimicking the behavior of the explainer model.

In the Pre-hoc Explainability Framework, for the ml-100k dataset, the accuracy values remain relatively stable across different λ_1 values, ranging from 0.7840 to 0.7740. On the other hand, the fidelity values gradually increase as λ_1 increases, starting from 0.8207 and reaching a peak at λ=1.0 with a fidelity value of 0.9410, which is a 14.6% increase. This suggests that higher λ_1 values in the Pre-hoc framework result in more faithful explanations without significantly sacrificing accuracy.

Similarly, for the ml-1M dataset, the accuracy values are relatively consistent, with the highest accuracy, 0.8076, observed at λ_1= 0.25. On the other hand, the fidelity values show an increase of at most 12. 3% when the value of λ_1 increases, ranging from 0.8769 to 0.9856. λ_1=1.0 achieves the highest fidelity, indicating that the Pre-hoc framework with a higher λ_1 value captures more accurate and informative explanations.

In the HELOC dataset, the accuracy values range from 0.7591 to 0.7719 across different λ_1 values. As λ_1 increases, the fidelity values also exhibit an upward trend, starting from 0.7482 and reaching a peak at λ=1.0 with a fidelity value of 0.8454. This suggests that the Pre-hoc framework with higher λ values enhances the fidelity of the explanations while maintaining comparable accuracy.

Similarly, in the Co-hoc Explainability Framework; for the ml-100k dataset, the accuracy values remain relatively stable, ranging from 0.7840 to 0.7766. The fidelity values show a gradual increase with increasing λ_1, starting from 0.8215 and reaching the highest fidelity of 0.9492 at λ_1=1. This suggests that the Co-hoc framework with a moderate λ value achieves better fidelity without compromising accuracy significantly.

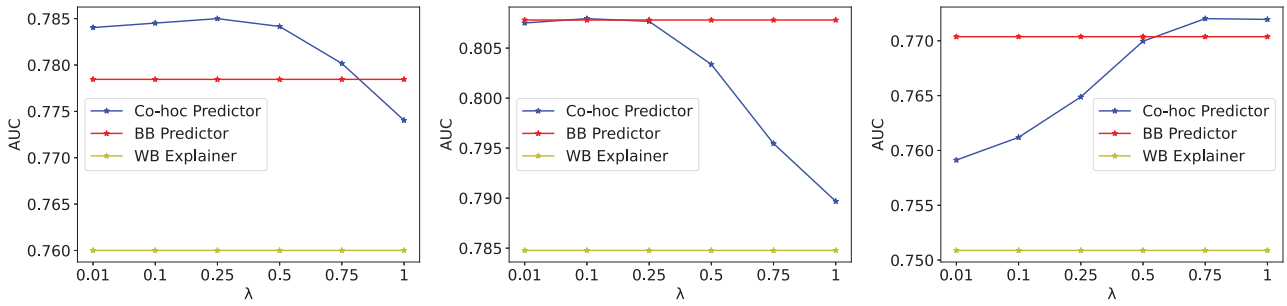
In the ml-1M dataset, the accuracy values are consistent across different λ_1 values, ranging from 0.8075 to 0.7924. On the other hand, the fidelity values show an increase of 12. 5%, starting at 0.8771 and reaching the highest fidelity of 0.9868 at λ_1=1. This indicates that the Co-hoc framework with higher λ values captures more faithful explanations while maintaining comparable accuracy.

In the HELOC dataset, the accuracy values range from 0.7591 to 0.7743 across different λ_1 values. As λ_1 increases, the fidelity values also exhibit an upward trend with a 14.5% increase, starting from 0.7482 and reaching the highest fidelity of 0.8572 at λ_1=1. This suggests that the Co-hoc framework with higher λ_1 values improves the fidelity of the explanations while maintaining similar accuracy levels.

Overall, the analysis of the experimental results shows that increasing λ_1 values lead to improvements in fidelity,

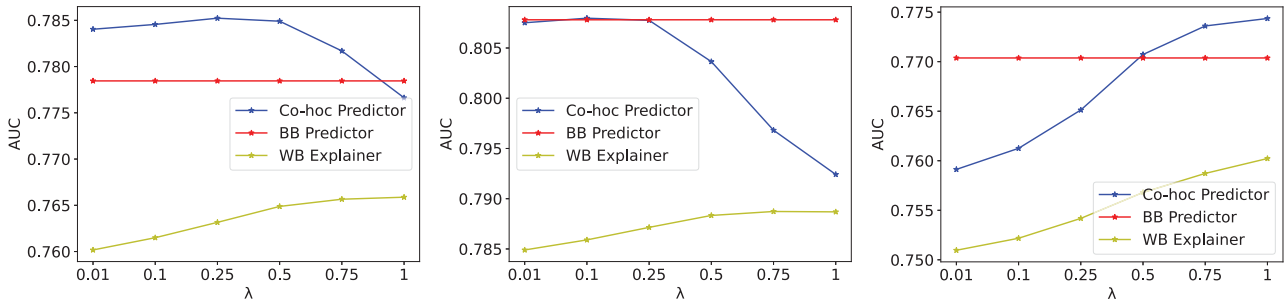
TABLE II: Proposed Explainability Frameworks, λ_1 comparison in prediction accuracy (AUC) and explainability (Fidelity) on the three datasets that were described in experimental settings. Higher AUC and Fidelity is better.

Framework	Dataset	Metric	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
Pre-hoc	ml-100k	AUC	0.7840	0.7845	0.7849	**0.7841**	0.7801	0.7740
		Fidelity	0.8207	0.8283	0.8430	0.8755	0.9094	**0.9410**
	ml-1M	AUC	0.8075	0.8079	**0.8076**	0.8033	0.7954	0.7896
		Fidelity	0.8769	0.8871	0.9058	0.9404	0.9696	**0.9856**
	HELOC	AUC	0.7591	0.7611	0.7648	0.7699	**0.7720**	0.7719
		Fidelity	0.7482	0.7541	0.7664	0.7903	0.8137	**0.8454**
Co-hoc	ml-100k	AUC	0.7840	0.7845	0.7852	**0.7849**	0.7816	0.7766
		Fidelity	0.8215	0.8326	0.8507	0.8869	0.9194	**0.9492**
	ml-1M	AUC	0.8075	0.8079	**0.8077**	0.8036	0.7968	0.7924
		Fidelity	0.8771	0.8901	0.9122	0.9484	0.9749	**0.9868**
	HELOC	AUC	0.7591	0.7612	0.7651	0.7707	0.7736	**0.7743**
		Fidelity	0.7482	0.7563	0.7705	0.7767	0.8277	**0.8572**



(a) (b) (c)

Fig. 3: Pre-hoc Explainability Framework Comparison in Accuracy AUC for different lambda on the ml-100k (a), ml-1M (b), HELOC (c) datasets. Pre-hoc Predictor is our proposed model, BB is the original black-box predictor model, WB is the explainer model.



(a) (b) (c)

Fig. 4: Co-hoc Explainability Framework Comparison in Accuracy AUC for different lambda on the ml-100k (a), ml-1M (b), HELOC (c) datasets. Co-hoc Predictor is our proposed model, BB is the original black-box predictor model, and WB is the explainer model.

indicating more accurate and faithful explanations with accuracy values remaining relatively stable or showing marginal variations across different λ_1 values. The optimal balance between accuracy and fidelity may vary depending on the dataset and the specific requirements of the explainability framework.

RQ4: What are the differences between the pre-hoc and co-hoc frameworks?

For comparing the two proposed frameworks, we consider accuracy, fidelity, and lambda sensitivity properties aspects.

Accuracy: In terms of prediction accuracy, both frameworks generally perform similarly across the evaluated datasets. In most cases, the accuracy values are comparable, with only slight variations observed. For example, in the ml-100k dataset, both frameworks achieve accuracy values around 0.78. Similarly, in the ml-1M and HELOC datasets, both frameworks achieve similar accuracy values with slight differences.

Fidelity: The fidelity of the Co-hoc framework tends to be consistently higher compared to the Pre-hoc framework. Across all datasets, the Co-hoc framework achieves higher fidelity scores, indicating that it better approximates the behavior of the explainer model. For instance, in the ml-100k dataset, the fidelity of Co-hoc ranges from 0.8215 to 0.9492, whereas Pre-hoc ranges from 0.8207 to 0.9410. The same trend applies to other datasets as well, see Table II.

Lambda Sensitivity: Both frameworks exhibit sensitivity to the choice of λ. The performance in terms of accuracy and fidelity can vary depending on the specific value of λ used. The optimal value of λ that maximizes the trade-off between fidelity and accuracy may differ between the two frameworks and across different datasets.

Overall, the Co-hoc Explainability Framework consistently demonstrates higher fidelity than the Pre-hoc Framework, while the differences in prediction accuracy between the two frameworks are relatively minor. This suggests that the Co-hoc approach, which jointly optimizes both the predictor model and explainer model, has the potential to approximate the mechanisms of the original model better and provide more accurate explanations. However, further analysis and experimentation are needed to fully understand the underlying factors contributing to these differences and their implications

in various contexts. One advantage of the co-hoc framework over the pre-hoc framework is that it has a more accurate white-box model due to the joint training phase.

VII. Conclusion

We proposed two novel approaches called Pre-hoc explainability and Co-hoc explainability frameworks that guide black-box predictor model training via an interpretable white-box model to align the black-box predictor's global logic with the white-box explainer's transparent reasoning rather than extracting post-hoc approximations of the white-box's logic. The proposed models incorporate the fidelity for any differentiable machine learning model without modifying the model architecture. Our work addresses the lack of explainability optimization during training and model-agnostic methods to enhance global explainability. Our future work will include extending our framework to produce local explanations.

The transparency of the white-box model may depend on the quality and quantity of the training data, as well as the complexity and heterogeneity of the underlying distribution. In particular, if the data are noisy or biased, or if the true relationship between the input and output variables is highly nonlinear or ambiguous. Thus, the white-box explainer model in our proposed framework could be easily replaced by alternative differentiable white-box models, such as rule-based or sparse additive models.

Additionally, our proposed approach does not explicitly address the issue of fairness or bias in the black-box model, which may be exacerbated by using pre-hoc and co-hoc explainability. Therefore, future work could explore how to incorporate fairness and bias considerations into our approach, develop complementary methods to mitigate these issues, and experiment with diverse datasets.

Acknowledgment

This work was supported by NSF-EPSCoR–RII Track-1:Kentucky Advanced Manufacturing Partnership for Enhanced Robotics and Structures (Award IIP#1849213) and by NSF DRL-2026584.

References

[1] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018.

[2] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2020.

[3] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," 2018.

[4] M. T. Ribeiro, S. Singh, and C. 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, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, oct 2019.

[7] S. Bordt, M. Finck, E. Raidl, and U. von Luxburg, "Post-hoc explanations fail to achieve their purpose in adversarial contexts," in *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022.

[8] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186.

[9] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *CoRR*, vol. abs/1806.07538, 2018. [Online]. Available: http://arxiv.org/abs/1806.07538

[10] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," 2019.

[12] G. Plumb, D. Molitor, and A. Talwalkar, "Model agnostic supervised local explanations," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 2520–2529.

[13] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 2801–2807.

[14] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[15] Z. C. Lipton, "The mythos of model interpretability," 2017.

[16] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," 2018.

[17] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2662–2670.

[18] G.-H. Lee, D. Alvarez-Melis, and T. S. Jaakkola, "Game-theoretic interpretability for temporal modeling," 2018.

[19] G.-H. Lee, W. Jin, D. Alvarez-Melis, and T. Jaakkola, "Functional transparency for structured data: a game-theoretic approach," in *International Conference on Machine Learning*, 2019.

[20] G. Plumb, M. Al-Shedivat, A. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 10 526–10 536.

[21] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, "A framework for learning ante-hoc explainable models via concepts," 2021.

[22] F. Nielsen and S. Boltz, "The burbea-rao and bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.

[23] T. M. Cover and J. A. Thomas, "Information theory and statistics," *Elements of information theory*, vol. 1, no. 1, pp. 279–335, 1991.

[24] F. Nielsen, "On the jensen–shannon symmetrization of distances relying on abstract means," *Entropy*, vol. 21, no. 5, p. 485, may 2019.

[25] GroupLens, "Movielens 100k dataset," https://grouplens.org/datasets/movielens/100k/.

[26] FICO, "The fico heloc dataset," https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2.

[27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, oct 2019.

[28] S. Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 995–1000.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.