Fidelity as a Scalar Representation for Post-Hoc XAI: Toward Uncertainty Quantification and Adversarial Detection

Yuchi Tang, Iñaki Esnaola, George Panoutsos
School of Electrical and Electronic Engineering, *The University of Sheffield*, Sheffield, United Kingdom
{ytang87, esnaola, g.panoutsos}@sheffield.ac.uk

***Abstract*—Post-hoc explainable AI (XAI) methods, particularly those that produce explanations formed by feature-wise importance scores, have become central to interpreting opaque AI models. However, post-hoc explanations often suffer from instability arising from their inherent stochasticity and are also vulnerable to adversarial attacks. In this paper, we introduce the FISCAR (FIdelity as a SCAlar Representation) framework, which transforms feature-level importance scores within a given post-hoc explanation into a scalar quantity. Subsequently, by modeling this scalar as a random variable, FISCAR provides an anchor for quantitatively analyzing explanation processes in relation to the inherent randomness within post-hoc XAI. We develop two methods based on this FISCAR framework: a Bayesian quantification method that uses the inferred inverse gamma distribution of the random variable to measure uncertainty, and a detection method that identifies adversarial behavior by monitoring the empirical variance of the random variable. Simulations on real-world datasets show that FISCAR provides a practical and effective entry point for these downstream tasks, which would otherwise lack a numerical foundation for further analysis. FISCAR thus enables more accountable evaluation of post-hoc explainability and supports the development of more trustworthy AI systems.**

***Index Terms*—Adversarial detection, explainable AI (XAI), explainability, post-hoc explanation, uncertainty**

I. Introduction

State-of-the-art AI models, particularly those based on machine learning and powered by large-scale data, are often regarded as opaque. This is because their internal structures are either inaccessible or too complex to be readily interpreted by humans. To foster trust in such AI models, a growing body of explainable AI (XAI) methods has emerged to help interpret how specific model outputs are generated, as summarized in [1], [2]. This need is critical in post-hoc scenarios, where predictions have already been made by a model, yet the rationale behind its decision making remains unclear.

Among existing approaches, post-hoc XAI methods that assign feature-wise importance scores have gained widespread adoption, particularly those that do not require access to internal model details (e.g., the exact neuron weights in deep neural networks). Representative examples include LIME [3] and SHAP [4], through which the generated feature scores allow individual assessment of each feature's contribution to the prediction, helping users interpret how different input components influence the model's decision. Model-specific methods, such as TreeSHAP [5] and Integrated Gradients [6], rely on access to the model's internal structure, which limits their portability.

Regardless of model accessibility, post-hoc XAI methods in general have been widely criticized for producing unreliable explanations [7]. Specifically, post-hoc explanations are often unstable [8] and vulnerable to adversarial manipulation [9], even without altering the instance being explained. This is because the inherent stochasticity of many post-hoc explanation methods causes their outputs to vary across repeated runs. Such variability not only undermines the reliability of explanations, but also creates opportunities for adversarial attacks to be concealed within this randomness, making them more difficult to detect. Especially in the more commonly encountered model-agnostic settings, these issues tend to be more pronounced, since the lack of access to the model's internal structure results in greater opacity.

This concern becomes even more significant as recent studies [10], [11] have demonstrated the use of post-hoc explanations to guide actionable input modifications that steer model predictions toward desired outcomes. In such cases, the stability of explanations directly affects the reliability of these interventions. When explanations themselves are unstable or susceptible to manipulation, the resulting interventions may become unreliable or misleading. This highlights the need to develop robust mechanisms for evaluating their uncertainty and detecting adversarial behavior.

In parallel with these developments, several fidelity metrics [12], [13], [14] have been proposed to assess how well post-hoc explanations reflect the actual behavior of the underlying model. These metrics quantify the extent to which the feature importance indicated by an explanation corresponds to the model's decision-making process. Notably, such metrics typically produce a single numerical value that maps a multi-dimensional explanation into a one-dimensional scalar. This scalar serves as a compressed representation of the explanation and can vary across different input instances. When considered alongside the inherent variability of post-hoc explanations, it offers a promising opportunity to observe explanation-level randomness.

In this paper, we present the following contributions:

- We propose the **FISCAR** (**FI**delity as a **SCA**lar **R**epresentation) framework, which transforms a given post-hoc explanation into a one-dimensional scalar. The

resulting scalar value is modeled as a random variable, allowing for numerical analysis of these explanations. This framework can be applied to a broad range of post-hoc XAI methods that produce feature-wise importance scores, without assuming any specific algorithmic structure.
- A FISCAR-based uncertainty quantification method; this employs Bayesian inference to quantify the uncertainty degree within a given post-hoc explanation through the derived posterior inverse gamma distribution of the FISCAR random variable.
- A FISCAR-based adversarial detection method; this method detects adversarial attacks that attempt to disguise biased models as innocuous ones by monitoring abnormal fluctuations in the FISCAR random variable.
- A validation of the proposed methods through numerical simulations on several real-world datasets. The results demonstrate that the proposed FISCAR framework is effective in uncertainty quantification and adversary detection for explanations produced by post-hoc XAI methods.

The code is available at: https://github.com/Yuchi-TANG-Research/FISCAR.

II. Preliminaries

A. Explanation Formed by Feature Importance Scores

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space for some $d \in \mathbb{N}$, and let $\mathcal{Y} \subseteq \mathbb{R}$ denote the output space. Define the feature index set as $G = \{1, \ldots, d\}$, corresponding to the components of an input $\mathbf{x} = (x_1, \ldots, x_d)$. To support understanding the prediction of an opaque model $f : \mathcal{X} \to \mathcal{Y}$ on a specific input-output pair $(\mathbf{x}, f(\mathbf{x}))$, an explanation formed by a vector of *feature important scores* $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d) \in \mathbb{R}^d$ can be produced. Specifically, each component score ϕ_i quantifies the individual importance of the corresponding input feature x_i to the specific model output $f(\mathbf{x})$.

B. Explaining Background

Let $P_{\mathbf{X}}^{(B)}$ denote a distribution, i.e., an explaining background distribution, defined over the input space $\mathcal{X}$, which is commonly adopted in existing post-hoc model-agnostic explanation methods to facilitate the computation of feature importance scores. Since its analytical form is typically unavailable in practice, $P_{\mathbf{X}}^{(B)}$ is approximated using a finite set of samples, i.e., explaining background samples $\{\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(c)}\}$ with $\mathbf{b}^{(i)} \in \mathcal{X}$ for $i = 1, \ldots, c$, where $c \in \mathbb{N}^+$.

C. Masked Output

Given an input $\mathbf{x}$ and a model f, together with a feature subset $S \subseteq G$, a *masked output* $\nu_f^S(\mathbf{x})$ is defined as the model's conditional prediction when the features in S are used, while the effect of other features are eliminated:

$$\nu_f^S(\mathbf{x}) = \mathbb{E}_{\mathbf{Z} \sim P_{\mathbf{X}}^{(B)}} \left[f(\mathbf{Z}) \,|\, \{Z_i = x_i\}_{i \in S} \right], \tag{1}$$

where $\mathbf{Z}$ is an auxiliary random variable that conditions on the features in S, allowing the influence of the remaining features to be removed from the output.

III. Problem Formulation

A. Existing post-hoc model-agnostic methods to provide feature importance scores

There are two primary approaches to producing feature importance scores in a post-hoc, model-agnostic setting: surrogate-based methods and attribution-based methods.

Surrogate-based methods explain the original opaque model by fitting an interpretable surrogate of it, which functions similarly as the original model within a defined input neighborhood. The feature importance scores can be thereby derived analytically from the surrogate model's interpretable structure. A prominent, widely recognized option of this approach is Local Interpretable Model-Agnostic Explanations (LIME) [3]. As a dominant local surrogate method, LIME builds an interpretable surrogate to approximate the prediction behavior of the opaque model f in the neighborhood of a given input $\mathbf{x}$. In most cases, this surrogate is set to be a linear model:

$$g(\mathbf{x}) = \sum_{i \in G} \phi_i x_i \approx f(\mathbf{x}), \tag{2}$$

where the coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$, i.e., the feature importance scores, are obtained by solving the following optimization to ensure similarity between the two models:

$$\underset{\text{(LIME)}}{\boldsymbol{\phi}} = \arg\min_{g \in \mathcal{G}} \ \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g), \tag{3}$$

where $\mathcal{G}$ is the set of all linear models parameterized by $(\phi_1, \ldots, \phi_d) \in \mathbb{R}^d$, and $\pi_{\mathbf{x}}$ defines the neighborhood of the the input $\mathbf{x}$, which typically relies on the selection of the samples in the explaining background $\mathcal{B}$, and $\Omega(g)$ is a regularization term that governs the complexity of g.

Attribution-based methods assign feature important scores to each feature by marginalizing out and isolating the feature's individual contribution to the output. The Shapley value [15], grounded in cooperative game theory, provides an axiomatic framework for systematically attributing feature importance scores. It gave rise to a wide range of attribution methods [4], [16], [17]. Among them, SHapley Additive exPlanations (SHAP) [4], has emerged as a prominent option. SHAP computes the importance score for each feature $i \in G$ as follows:

$$\underset{\text{(SHAP)}}{\phi_i(f, \mathbf{x})} = \sum_{S \subseteq G \setminus \{i\}} w_{\text{Shapley}}(S) \left[\nu_f^{S \cup \{i\}}(\mathbf{x}) - \nu_f^S(\mathbf{x}) \right], \tag{4}$$

where $w_{\text{Shapley}}(S) = |S|!(|G| - |S| - 1)! \,/\, |G|!$ are the Shapley weights that are determined by the cardinality of features in S and G.

A wide range of existing post-hoc model-agnostic XAI methods [18], [19], [20], [21] that produce feature-wise importance scores can be viewed as superseded preliminary versions or improved extensions of the methods discussed above.

B. Factors Leading to Randomness in Post-Hoc Explanations

Given a to-be-explained instance $(\mathbf{x}, f(\mathbf{x}))$, post-hoc model-agnostic methods may still yield varying feature importance scores, even when the same parameter settings are used,

as demonstrated in [22], [23]. This randomness arises from the inherently stochastic nature of the explaining process. Specifically, there are two primary factors contributing to such randomness: (F1) the stochastic sampling process used to select data points that constitute the explaining background, and (F2) the approximation strategies adopted to accelerate these otherwise computationally intensive methods.

Factor (F1) introduces randomness through variation in the choice of the explaining background distribution $P_{\mathbf{X}}^{(B)}$. As a result, both $\pi_{\mathbf{x}}$ in (3) for LIME and $\nu_f^S(\mathbf{x})$ in (4) for SHAP fluctuate as the background changes.

Factor (F2) introduces randomness through approximation strategies used to accelerate computation. For example, SHAP employs random sampling over feature permutations to approximate the exhaustive subset enumeration for $S \subseteq G$, particularly when the feature space $|G|$ is large. Similarly, LIME relies on generating randomly perturbed instances to fit the local surrogate model.

IV. Proposed Framework and Instantiated Methods

We propose the FISCAR framework to transform post-hoc explanations formed by feature importance scores into fidelity-based one-dimensional scalars. Building on this framework, we further instantiate its use in methods for uncertainty quantification and adversarial detection.

A. Proposed FISCAR Framework: FIdelity as a SCAlar Representation

Given an explanation formed by feature importance scores $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$, we define the feature importance rank of each score ϕ_i for $i \in G$, based on its absolute magnitude relative to the other scores. The ranking function $\rho : \mathbb{R} \to \{1, \ldots, d\}$ satisfies the following conditions:

- If $|\phi_i| > |\phi_j|$ for $i, j \in G$, then $\rho(\phi_i) < \rho(\phi_j)$.
- Let $\Phi^{i-} = \phi_p, |\phi_p| < |\phi_i|$ for $p \in G$, and let $|\phi_k| = \max |\Phi^{i-}|$, then $\rho(\phi_k) = \rho(\phi_i) + 1$.
- If $|\phi_i| = \max |\phi_1|, \ldots, |\phi_d|$, then $\rho(\phi_i) = 1$.

This ranking-based formulation enables the assessment of explanation uncertainty by quantifying how sensitive feature importance rankings are to the inherent randomness of post-hoc methods.

Based on this, we introduce a scalar representation of feature importance scores, rather than directly relying on the multi-dimensional importance values. Motivated by the fidelity metric known as the Area Under the Prediction recovery curve (AUP) [20], we introduce its variant, Squared AUP (SAUP), as a compact one-dimensional value that summarizes $\boldsymbol{\phi}$.

Specifically, let $\mathcal{I}(\rho, m; \boldsymbol{\phi}) \subseteq [d]$ be the set of indices that indicate the m most important features with respect to the ranking vector $\boldsymbol{\rho} = (\rho(\phi_1), \ldots, \rho(\phi_d))$. With $\mathcal{I}(\rho, m; \boldsymbol{\phi})$, we have SAUP as follows:

$$\text{SAUP}(\boldsymbol{\rho}; \mathbf{x}, f) = \sum_{m=1}^{d} \left[f(\mathbf{x}) - \nu_f^{\mathcal{I}(\rho, m; \boldsymbol{\phi})}(\mathbf{x}) \right]^2. \tag{5}$$

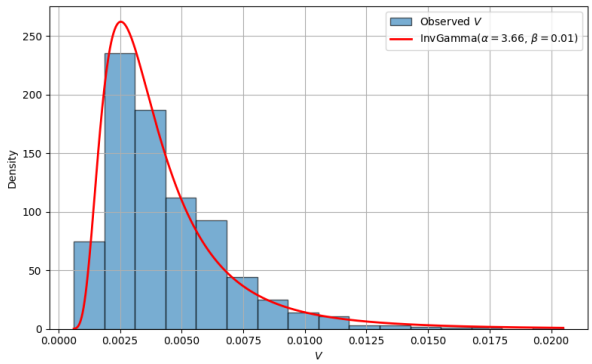

Fig. 1. Empirical distribution of V that resembles the shape of an inverse gamma distribution. These values are obtained by repeated runs of SHAP on an instance from the *concrete* dataset

Notably, SAUP consolidates the effects of all features within an explanation, accounting for each feature's influence in a collective scalar representation.

Given the inherent randomness caused by (F1) and (F2), we treat the SAUP value as a random variable V. Since SAUP aggregates squared terms, it satisfies $V \geq 0$. Moreover, SAUP can be interpreted as a proxy for a local variance-like quantity capturing the dispersion of the masked outputs $\nu_f^{\mathcal{I}(\rho, m; \boldsymbol{\phi})}(\mathbf{x})$. In classical statistical settings, variance parameters associated with Gaussian observations are often modeled by inverse gamma distributions. As in [24], when residuals are Gaussian and the variance parameter is unknown, Bayesian formulations commonly yield an inverse gamma distribution as a conjugate prior for the variance. Motivated by this, we approximately model the masked outputs as i.i.d. Gaussian variables. In view of this, the squared aggregation suggests that V can be reasonably approximated in distribution by an inverse gamma distribution:

$$V \sim \text{InvGamma}(\alpha, \beta), \tag{6}$$

where $\alpha, \beta \in \mathbb{R}^+$ denote the shape and scale parameters. Empirically, as shown in Fig. 1, repeated SHAP runs on a given instance produce an empirical distribution of V that closely resembles the characteristic shape of an inverse gamma distribution.

B. FISCAR-Based Uncertainty Quantification

With the FISCAR framework, a post-hoc explanation is considered as a realization of a random variable V, which is modeled to approximately follow an inverse gamma distribution. This probabilistic perspective enables the estimation of distributional parameters, thereby allowing the uncertainty associated with the explanation to be quantitatively assessed.

To achieve this, we adopt a hierarchical Bayesian framework in which hyperpriors are placed on the shape and scale parameters α and β of the inverse gamma distribution. This facilitates a principled estimation of the distribution of V:

$$\alpha \sim \text{Gamma}(\gamma_0, \delta_0), \tag{7}$$

$$\beta \sim \text{Gamma}(\lambda_0, \theta_0), \tag{8}$$

where $\gamma_0, \delta_0, \lambda_0, \theta_0 \in \mathbb{R}^+$. We set $\gamma_0 = \lambda_0 = 1$ and $\delta_0 = \theta_0 = 0.1$, corresponding to weakly informative Gamma priors commonly used in Bayesian modeling. The formulations in (7)(8) enable propagating the uncertainty of the distributional parameters into the posterior estimation of V. To this end, we explicitly model the likelihood of an observed value v given α and β as follows:

$$p(v \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} v^{-(\alpha+1)} \exp\left(-\frac{\beta}{v}\right), \quad v > 0 \tag{9}$$

Given the likelihood in (9) and the hyperpriors in (7)(8), we aim to infer the joint posterior distribution over the shape and scale parameters:

$$p(\alpha, \beta \mid v) \propto p(v \mid \alpha, \beta) \cdot p(\alpha) \cdot p(\beta). \tag{10}$$

Since this posterior does not admit a closed-form solution, we employ Markov chain Monte Carlo (MCMC) sampling to draw posterior samples from the joint distribution of α and β to finalize their estimations. In particular, we adopt the No-U-Turn Sampler (NUTS), a gradient-based variant of Hamiltonian Monte Carlo (HMC), to approximate the joint posterior distribution $p(\alpha, \beta \mid \{v_n\})$, where $\{v_n\}$ denotes repeated SAUP observations obtained by re-running the explanation procedure on the same input. NUTS enables efficient exploration of the posterior space without requiring manual tuning of proposal distributions or conjugacy assumptions.

In our implementation, we run four independent Markov chains, each with 500 warm-up iterations followed by 1000 post-warm-up samples, resulting in a total of 4000 posterior samples. Given posterior samples $\{(\alpha^{(k)}, \beta^{(k)})\}_{k=1}^{K}$, we construct a posterior predictive distribution by sampling $V^{(k)} \sim \mathrm{InvGamma}\big(\alpha^{(k)}, \beta^{(k)}\big)$.

Uncertainty quantification is then performed directly on the posterior predictive samples $\{V^{(k)}\}$. Specifically, we use the posterior predictive V's 95% highest density interval (HDI) for the quantification. This yields a statistically grounded characterization of the uncertainty associated with the explanation under repeated instantiations of the post-hoc explanation procedure.

C. FISCAR-Based Adversarial Detection

Post-hoc XAI methods, especially in a model-agnostic setting, have been criticized for their unreliability in the absence of ground-truth information [25], [26]. In this context, adversarial attacks can further expose their vulnerability. In [27], it has been demonstrated how a biased classifier can be constructed while concealing its discriminatory behavior, whereas such adversary should have been revealed through abnormal feature importance scores on the attacked sensitive features.

Specifically, given an unbiased model ψ (i.e., predicts from innocuous features) and a biased model θ (i.e., predicts from sensitive features), the adversarial classifier ϵ is formulated as follows:

$$\epsilon(\mathbf{x}) = \begin{cases} \theta(\mathbf{x}), & \text{if } \mathbf{x} \in \mathcal{X}_{dist} \\ \psi(\mathbf{x}), & \text{otherwise} \end{cases}, \tag{11}$$

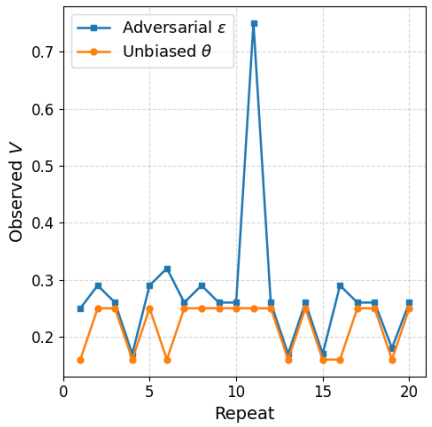

Fig. 2. Observed V values with unbiased and adversarial models with 20 repeated SHAP explanations for a single instance from the *COMPAS* dataset. Compared to the unbiased model, the adversarial model tends to exhibit higher variance in the observed V value.

where $\mathcal{X}_{dist}$ denotes the true input distribution of the dataset. By leveraging an out-of-distribution (OOD) discriminator to determine whether $\mathbf{x} \in \mathcal{X}_{dist}$, the adversarial classifier produces biased outputs while simultaneously presenting seemingly normal explanations. This occurs because OOD inputs can be artificially generated and subsequently detected during the explanation process. As a result, the abnormal patterns that post-hoc methods would otherwise be expected to reveal are effectively concealed, as shown in [27].

Still, such adversarial interventions induce irregular fluctuations in the values of $\nu_f^{\mathcal{I}(\rho,m;\boldsymbol{\phi})}(\mathbf{x})$, which in turn lead to anomalous deviations in the distribution of V. As illustrated in Fig. 2, the adversarial model tends to exhibit larger empirical variance in the resulting FISCAR values V in practice.

Therefore, under the proposed FISCAR framework, the scalar representation V provides a valuable signal for detecting adversarial manipulation in post-hoc explanations. By monitoring the abnormal variability of V across repeated explanation generations, namely $\widehat{\mathrm{Var}}(V)$, it becomes possible to flag suspicious behaviors that may indicate tampering or instability. This variance-based signal supports the recognition of such behaviors without relying on model-specific assumptions, thereby facilitating adversary detection.

V. Empirical Evaluations

A. Simulation Setups

We design the following numerical simulations with the following setups to demonstrate the applicability of the proposed methods in uncertainty quantification and adversarial detection, respectively.

For uncertainty quantification, we conduct simulations on publicly available regression datasets *abalone* [28], *california* [29], and *concrete* [30], using XGBoost models [31] trained on each. We artificially introduce variations into post-hoc explanations and examine whether the resulting uncertainty measures fluctuate accordingly:

TABLE I
MEAN DIFFERENCE OF 95% HDI WIDTHS BETWEEN HIGH- AND LOW-UNCERTAINTY SETTINGS ACROSS TEST INSTANCES.

Dataset	Mean difference with 95% CI ($\times 10^{-2}$)	
	LIME	SHAP
Abalone	3.173 (2.587, 3.903)	0.419 (0.261, 0.597)
California	9.407 (7.878, 11.029)	0.097 (0.056, 0.144)
Concrete	15.207 (12.643, 17.821)	0.234 (0.158, 0.316)

TABLE II
MEAN DIFFERENCE OF EMPIRICAL VARIANCE OF V BETWEEN ADVERSARIAL AND UNBIASED MODELS ACROSS TEST INSTANCES.

Dataset	Mean difference with 95% CI	
	LIME	SHAP
COMPAS	0.499 (0.466, 0.528)	0.075 (0.068, 0.083)
German credit	6.053 (4.109, 8.296)	0.109 (0.047, 0.212)

- With LIME explanations, we vary the uncertainty level by adjusting the proportion of background data actively used in the explanation process. To simulate a higher uncertainty level, we sample only 10% of the available background data. Conversely, a lower uncertainty level is simulated by using 100% of the background data.
- With SHAP explanations, rather than enumerating all $S \subseteq G$ as in (4), we simulate different uncertainty levels by adjusting the approximation degree through the sampling ratio of S. This setup aligns with the internal design of SHAP's `PermutationExplainer`, where the sampling ratio is configurable. Specifically, a lower sampling ratio of 20% is used to induce a higher uncertainty level in the explanations, whereas a higher ratio of 80% corresponds to a lower uncertainty level.

For adversarial detection, we adopt the same adversarial configurations and decision-making models on the *COMPAS* [32] and *German Credit* [33] datasets as used in [27]. We compute the empirical variance $\widehat{\mathrm{Var}}(V)$ for each test instance under both the adversarial model ϵ and the unbiased model ψ by repeating the SHAP and LIME explanation processes 20 times. To further evaluate the capability of FISCAR-based detection, we instantiate a prototype method by learning an optimal threshold τ on a randomly selected 10% of the test set. An instance is flagged as attacked if its $\widehat{\mathrm{Var}}(V)$ exceeds τ. This threshold is chosen to maximize detection performance and is subsequently evaluated on the remaining 90% of the test set. For comparison, we also implement the detection method from [27] on the same evaluation instances. In that method, an instance is classified as adversarial if the manipulated sensitive feature (i.e., *race* in COMPAS and *Gender* in German Credit), which should not influence the model's decision, appears among the top-k ($k \in [1, 2, 3]$) most important features in the explanation output.

B. Results and analysis

The FISCAR-based quantification method consistently aligns with the varied uncertainty levels.

As shown in Table I, the FISCAR-based uncertainty quantification method consistently yields positive mean differences in the widths of the 95% HDIs between explanations generated under high- and low-uncertainty settings. Here, the HDIs are estimated empirically from 100 repeated explanation runs for each instance. For each hold-out test instance, we compute the HDI width under the high-uncertainty setting and subtract the width under the low-uncertainty setting. A positive result indicates that the method assigns broader widths when the explanation process is more stochastic. These per-instance differences are then aggregated across 100 bootstrap samples to estimate the mean and corresponding confidence intervals. The reported positive means, with no bootstrap intervals crossing zero, show that higher simulated uncertainty consistently leads to broader HDIs. This reflects increased posterior variability and confirms that the method effectively captures and quantifies uncertainty in response to changes in explanation-generating conditions across all datasets.

At the instance level, Fig. 3 presents posterior distributions inferred via the full Bayesian inference procedure of the proposed approach, as described in Section IV-B, under low-uncertainty and high-uncertainty settings. These distributions are constructed from repeated explanation outcomes generated by LIME and SHAP across individual input instances. Under high-uncertainty conditions, the inferred inverse gamma posteriors consistently exhibit heavier tails and wider 95% HDIs, indicating greater variability in the scalar values V constructed under the FISCAR framework. In contrast, the low-uncertainty settings yield sharper distributions with narrower HDIs. This instance-level behavior supports the method's ability to capture explanation uncertainty induced by the inherent randomness of post-hoc XAI methods.

Furthermore, **the FISCAR-based detection method reveals clear behavior differences between adversarial and unbiased models.**

As shown in Table II, across both datasets and explanation methods, the estimated mean differences in variance, defined as the variance under adversarial models minus that under unbiased models, are consistently positive, and the corresponding 95% confidence intervals do not include zero. This indicates that adversarial models systematically induce higher empirical variance in V compared to their unbiased counterparts. The consistency of this pattern suggests that the observed variance differences reflect a stable characteristic of adversarial behavior under repeated post-hoc explanation. The uniformly positive variance differences demonstrate that increased variability under adversarial models is reliably captured by the variance of V under the proposed FISCAR framework, supporting its role as a stable and informative indicator of adversarial influence.

As shown in Table III, the performance of adversarial detection on the *COMPAS* and *German Credit* datasets demonstrates the effectiveness of the FISCAR-based approach with fitted detection thresholds τ. For both SHAP and LIME explanations, the learned thresholds achieve substantially higher

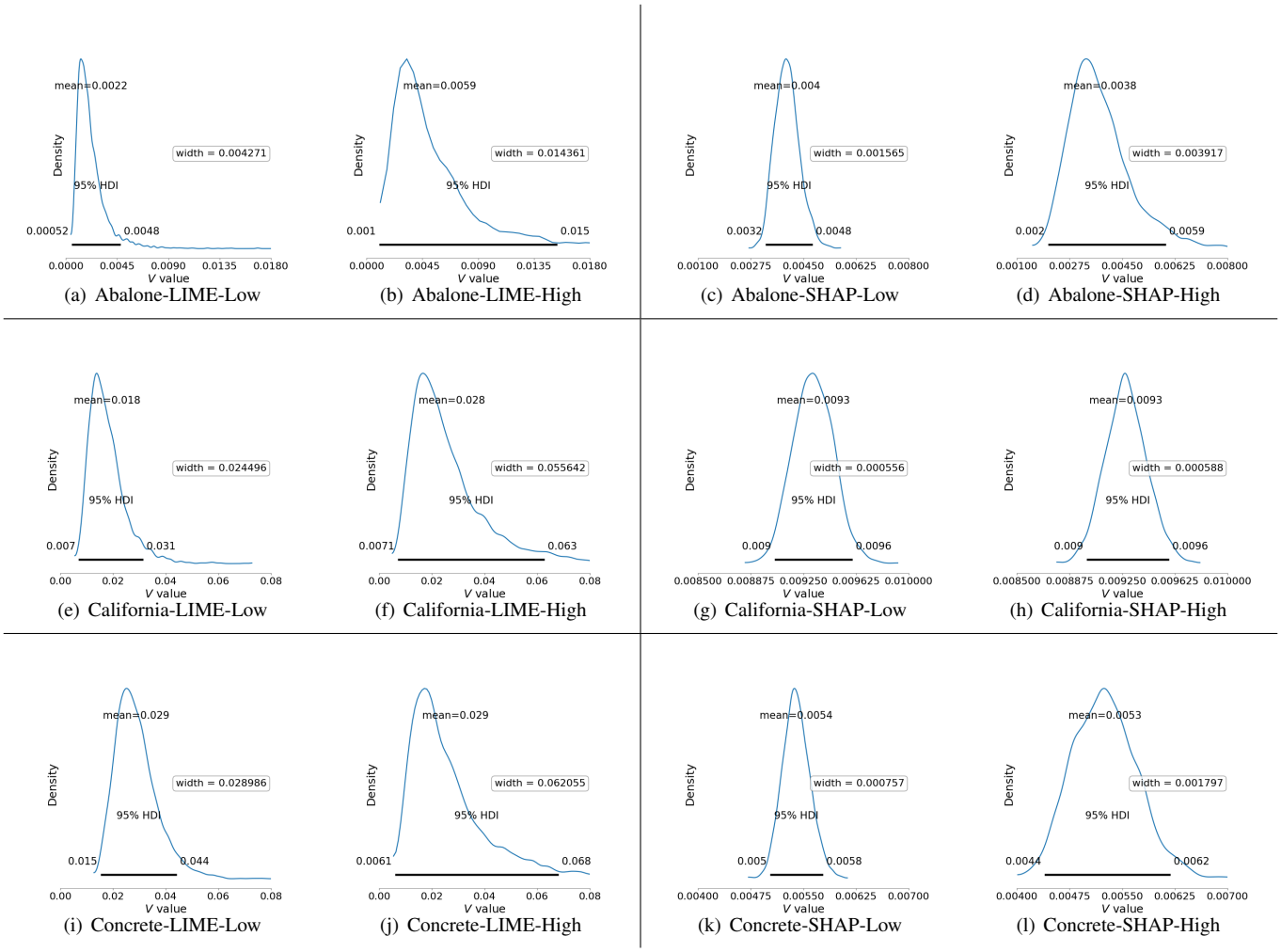

Fig. 3. Instance-level examples of LIME and SHAP explanations under simulated low and high uncertainty. For each pair of explanations for the same instance, the FISCAR-based method yields broader 95% HDIs under the high-uncertainty setting.

TABLE III
PERFORMANCE OF ADVERSARIAL DETECTION.

Dataset	**Strategy**		**Adversary detected (%)**	
			LIME	**SHAP**
COMPAS	Top-k occurrence check	$k=1$	0.72	19.60
		$k=2$	5.04	39.39
		$k=3$	14.57	70.50
	FISCAR-τ		**96.85**	**97.57**
German credit	Top-k occurrence check	$k=1$	0.00	10.00
		$k=2$	0.00	17.78
		$k=3$	2.22	25.56
	FISCAR-τ		**100.0**	**83.89**

detection rates compared to the baseline top-k occurrence check methods. On the *COMPAS* dataset, the FISCAR-based detector identifies over 96% of adversarial instances on both types of post-hoc explanations, while the top-k method detects no more than 70.50% of adversarial instances even when $k = 3$. The difference is even more pronounced on the *German Credit* dataset, where the FISCAR method reaches a 100.0% detection rate for LIME and 83.89% for SHAP. In contrast, the top-k baseline detects fewer than 26% of adversarial instances, even with the most generous k value. These results highlight the advantage of the proposed scalar-based detection strategy, which operates by tracking the empirical variance of V across repeated explanation generations. Unlike heuristic methods that rely on the ranking of individual features, the FISCAR framework offers a mathematically principled mechanism for identifying adversarial behavior in post-hoc XAI.

VI. Related Work

A. Uncertainty Quantification of Post-Hoc XAI

Efforts to quantify uncertainty in post-hoc XAI have resulted in a variety of techniques that are often tailored to specific explanation methods. In [34], the uncertainty caused by the stochastic sampling process in LIME is examined through the statistical behavior of the weighted Ridge regression used in its local surrogate model. In [35], each LIME coefficient is modeled within a Bayesian framework under Gaussian noise

assumptions and conjugate priors, allowing uncertainty to be expressed through posterior distributions. In [19], Gaussian modeling is applied to the subset sampling process used in SHAP to capture uncertainty in the generated importance scores. In [36], confidence intervals and hypothesis tests are constructed based on the assumption that the joint distribution of SHAP values is pseudo-elliptical. These methods are largely grounded in the internal algorithmic structure of their respective explanation techniques. As a result, their applicability to other methods remains limited. This limitation becomes more pronounced as new post-hoc XAI approaches continue to emerge.

In [8], [37], quantification methods are proposed that can be applied to any explanation technique producing feature-wise importance scores. Nonetheless, the resulting uncertainty measures are still computed independently for each feature within a given instance, as is also the case in [19], [34], [35], [36]. Their per-feature credible interval structures complicate the integration of uncertainty information into downstream tasks and may limit its usability for end users, particularly in high-dimensional settings. More importantly, such approaches fail to capture the joint behavior among feature importance scores. For instance, users cannot determine whether a large importance score for one feature coincides with high or low for others, even when all scores lie within their respective credible intervals. As a result, potential interactions and dependencies across features are systematically overlooked.

B. Adversarial Detection in Post-Hoc XAI

Post-hoc XAI systems are vulnerable to adversarial manipulation. Prior work [38] showed that explanations of gradient-based models can be substantially altered via imperceptible input perturbations that preserve the original model predictions. Beyond input-level attacks, in [39], models are explicitly trained to conceal their reliance on sensitive attributes by jointly optimizing predictive performance and explanation-based regularization, thereby misleading multiple post-hoc explanation methods. However, such attacks typically require privileged access to model internals or direct manipulation during training, making them inherently conspicuous. Without requiring such access, as demonstrated in [27], an attacker can deploy a biased model that activates only when a specific feature is present. Under such conditions, post-hoc model-agnostic explanation methods, including widely used LIME and SHAP, can still produce explanations that appear deceptively normal.

Since many adversarial approaches tend to target a specific explanation method, the work in [40] proposes a defense strategy that aggregates multiple explanation methods in order to provide more robust explainability. However, due to the shared and inherent randomness present in most post-hoc methods, particularly those that are model-agnostic, such aggregation remains susceptible to manipulation. A more targeted solution is proposed in [41], which introduces a constraint-based method for LIME to produce explanations that are more stable. In addition, the study in [42] explores the use of a more localized neighborhood of background data in SHAP to reduce the produced explanation's vulnerability. These approaches primarily operate by improving the explanation algorithms themselves. Therefore, they function as passive enhancement mechanisms and do not offer an analytical framework for actively and quantitatively detecting adversarial behavior. Nevertheless, adversarial detection for post-hoc explanations is still at an early stage, with few methods offering generalizable solutions. This limitation has been noted in the review [43], which concludes that "conceiving strategies to improve the reliability and robustness of explanation methods continues to be an urgent line of research."

VII. Conclusion

This paper presents the FISCAR framework, a fidelity-based approach for transforming post-hoc explanations into one-dimensional scalar values. By framing these scalars as random variables, we enable systematic, numerical analysis of post-hoc explanations formed by feature importance scores. Grounded in this framework, we propose a Bayesian method for uncertainty quantification and a variance-based method for adversary detection. Our numerical simulations, conducted on widely adopted LIME and SHAP explanations across multiple datasets, demonstrate the advantages of FISCAR over existing baselines. The scalar formulation captures the variability introduced by the inherent randomness of post-hoc XAI methods and supports further evaluations on explanation reliability. These findings suggest that FISCAR can serve as an effective foundation for downstream tasks that rely on post-hoc explanations, enabling them to operate on quantifiable signals.

Beyond the demonstrated results, the FISCAR framework opens promising directions for future research. Its design provides a flexible foundation that can be extended to support a wider range of use cases with XAI outcomes. Still, a limitation of the current framework is that the inverse-gamma modeling adopted for SAUP provides only an approximate probabilistic characterization, as potential correlations among masked outputs are not explicitly modeled. The uncertainty quantification method introduced in this work may therefore be further enhanced through richer probabilistic modeling or alternative inference strategies that more tightly integrate with SAUP or other scalar formulations derived from explanation mappings. Likewise, the adversarial detection method can be further refined to target-specific classes of attacks, enabling more precise and adaptive identification mechanisms.

Acknowledgments

This work was supported by the UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership (DTP) through the Healthy Lifespan Institute (HELSI) Flagship Scholarship at The University of Sheffield (Grant No. EP/W524360/1). The authors acknowledge the use of ChatGPT (OpenAI) for language improvements.

REFERENCES

[1] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artif Intell Rev*, vol. 55, no. 5, pp. 3503–3568, 2022.

[2] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *Acm Comput Surv*, vol. 55, no. 9, pp. 1–33, 2023.

[3] 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*, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[5] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, 2020.

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[7] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[8] C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon, and L. Huan, "But are you sure? an uncertainty-aware perspective on explainable ai," in *International conference on artificial intelligence and statistics*. PMLR, 2023, pp. 7375–7391.

[9] S. Bordt, M. Finck, E. Raidl, and U. Von Luxburg, "Post-hoc explanations fail to achieve their purpose in adversarial contexts," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 891–905.

[10] S. Gyawali, J. Huang, and Y. Jiang, "Leveraging explainable ai for actionable insights in iot intrusion detection," in *2024 19th Annual System of Systems Engineering Conference (SoSE)*. IEEE, 2024, pp. 92–97.

[11] W. Dossche, S. Vansteenkiste, B. Baesens, and W. Lemahieu, "Anticipating delays in recruitment: Explainable machine learning for the prediction of hard-to-fill online job vacancies," *Eur J Oper Res*, 2025.

[12] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, 2016.

[13] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/92650b2e92290c3e74fbf864e1558c0f-Abstract.html

[14] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," *arXiv preprint arXiv:2005.00631*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00631

[15] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton: Princeton University Press, 1953, pp. 307–317.

[16] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl Inf Syst*, vol. 41, pp. 647–665, 2014.

[17] P. Kolpaczki, V. Bengs, M. Muschalik, and E. Hüllermeier, "Approximating the shapley value without marginal contributions," in *Proceedings of the AAAI conference on Artificial Intelligence*, 2024, pp. 13 246–13 255.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[19] I. Covert and S.-I. Lee, "Improving kernelshap: Practical shapley value estimation using linear regression," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3457–3465.

[20] Y. Kwon and J. Y. Zou, "Weightedshap: Analyzing and improving shapley based feature attributions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 363–34 376, Dec. 2022.

[21] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6968–6972, 2022.

[22] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[23] J. Cohen, E. Byon, and X. Huan, "To trust or not: Towards efficient uncertainty quantification for stochastic shapley explanations," in *Phm society asia-pacific conference*, vol. 4, 2023.

[24] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, "How much should i trust you? modeling uncertainty of black box explanations," *ArXiv*, Aug. 2020.

[25] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International conference on machine learning*. PMLR, 2020, pp. 5491–5500.

[26] X. Huang and J. Marques-Silva, "On the failings of shapley values for explainability," *Int J Approx Reason*, vol. 171, p. 109112, 2024.

[27] 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*, 2020, pp. 180–186.

[28] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford, "Abalone," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C55C7W.

[29] scikit-learn developers, "California housing dataset," 2024, available at https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html.

[30] I.-C. Yeh, "Concrete Compressive Strength," UCI Machine Learning Repository, 1998, DOI: https://doi.org/10.24432/C5PK67.

[31] T. Chen, "Xgboost: A scalable tree boosting system," *Cornell University*, 2016.

[32] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," *ProPublica*, May 2016, accessed: 2016-05. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[33] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77.

[34] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, "Statistical stability indices for lime: Obtaining reliable explanations for machine learning models," *J Oper Res Soc*, vol. 73, no. 1, pp. 91–101, 2022.

[35] Y.-H. Hung and C.-Y. Lee, "Bmb-lime: Lime with modeling local nonlinearity and uncertainty in explainability," *Knowl-Based Syst*, vol. 294, p. 111732, 2024.

[36] D. Fryer, I. Strümke, and H. Nguyen, "Shapley value confidence intervals for attributing variance explained," *Front Appl Math Stat*, vol. 6, Dec. 2020.

[37] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju, "Reliable post hoc explanations: Modeling uncertainty in explainability," *Advances in neural information processing systems*, vol. 34, pp. 9391–9404, 2021.

[38] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[39] B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller, "You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods," in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, vol. 97. IOS Press, 2020, pp. 161–170.

[40] L. Rieger and L. K. Hansen, "A simple defense against adversarial attacks on heatmap explanations," in *Proceedings of the Workshop on Human Interpretability in Machine Learning*, 2020.

[41] A. A. Shrotri, N. Narodytska, A. Ignatiev, K. S. Meel, J. Marques-Silva, and M. Y. Vardi, "Constraint-driven explanations for black-box ml models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8304–8314.

[42] S. Ghalebikesabi, L. Ter-Minassian, K. Diaz-Ordaz, and C. C. Holmes, "On locality of local explanation models," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 18 395–18 407.

[43] J. Vadillo, R. Santana, and J. A. Lozano, "Adversarial attacks in explainable machine learning: A survey of threats against models and humans," *Wires Data Min Knowl*, vol. 15, no. 1, p. e1567, 2025.