EXPLANATION UNCERTAINTY WITH DECISION BOUNDARY AWARENESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc explanation methods have become increasingly depended upon for understanding black-box classifiers in high-stakes applications, precipitating a need for reliable explanations. While numerous explanation methods have been proposed, recent works have shown that many existing methods can be inconsistent or unstable. In addition, high-performing classifiers are often highly nonlinear and can exhibit complex behavior around the decision boundary, leading to brittle or misleading local explanations. Therefore, there is an impending need to quantify the uncertainty of such explanation methods in order to understand when explanations are trustworthy. We introduce a novel uncertainty quantification method parameterized by a Gaussian Process model, which combines the uncertainty approximation of existing methods with a novel geodesic-based similarity which captures the complexity of the target black-box decision boundary. The proposed framework is highly flexible—it can be used with any black-box classifier and feature attribution method to amortize uncertainty estimates for explanations. We show theoretically that our proposed geodesic-based kernel similarity increases with the complexity of the decision boundary. Empirical results on multiple tabular and image datasets show that our decision boundary-aware uncertainty estimate improves understanding of explanations as compared to existing methods.

1 INTRODUCTION

Machine learning models are becoming increasingly prevalent in a wide variety of industries and applications. In many such applications, the best performing model is opaque; post-hoc explainability methods are one of the crucial tools by which we understand and diagnose the model's predictions. Recently, many explainability methods, termed *explainers*, have been introduced in the category of local feature attribution methods. That is, methods that return a real-valued score for each feature of a given data sample, representing the feature's relative importance with respect to the sample prediction. These explanations are local in that each data sample may have a different explanation. Using local feature attribution methods therefore helps users better understand nonlinear and complex black-box models, since these models are not limited to using the same decision rules throughout the data distribution.

Recent works have shown that existing explainers can be inconsistent or unstable. For example, given similar samples, explainers might provide different explanations (Alvarez-Melis & Jaakkola, 2018; Slack et al., 2020). When working in high-stakes applications, it is imperative to provide the user with an understanding of whether an explanation is reliable, potentially problematic, or even misleading. A way to guide users regarding an explainer's reliability is to provide corresponding *uncertainty quantification* estimates.

One can consider explainers as function approximators; as such, standard techniques for quantifying the uncertainty of estimators can be utilized to quantify the uncertainty of explainers. This is the strategy utilized by existing methods for producing uncertainty estimates of explainers (Slack et al., 2021; Schwab & Karlen, 2019). However, we observe that for explainers, this is not sufficient; because in addition to uncertainty due to the function approximation of explainers, explainers also have to deal with the uncertainty due to the complexity of the decision boundary (DB) of the black-box model in the local region being explained.

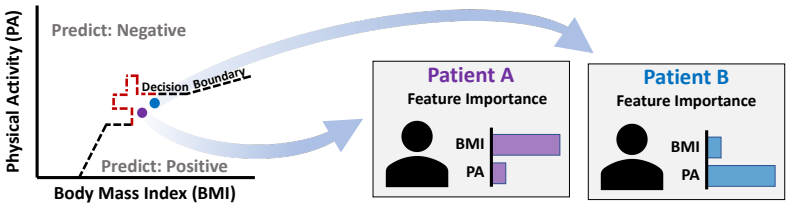

Figure 1: Illustrative example of how similar data samples can result in very different feature importance scores given a black-box model with nonlinear decision boundary. Here, two similar patients with similar predictions are given opposing feature importance scores which could result in misguided recommendations. We define a similarity based on the geometry of the decision boundary between any two given samples (red line). While the two patients are close together in the Euclidean sense, they are dissimilar under the proposed WEG kernel similarity. Using GPEC would return a high uncertainty measure for the explanations, which would flag the results for further investigation.

Consider the following example: we are using a prediction model for a medical diagnosis using two features, level of physical activity and body mass index (BMI) (Fig. 1). In order to understand the prediction and give actionable recommendations to the patient, we use a feature attribution method to evaluate the relative importance of each feature. Because of the nonlinearity of the prediction model, patients A and B show very similar symptoms, but are given very different recommendations. Note that while this issue is related to the notion of explainer uncertainty, measures of uncertainty that only consider the explainer would not capture this phenomenon. This suggests that any notion of uncertainty is incomplete without capturing information related to the local behavior of the model. Therefore, the ability to quantify uncertainty for DB-related explanation instability is desirable.

We approach this problem from the perspective of similarity: given two samples and their respective explanations, how closely related should the explanations be? From the previous intuition, we define this similarity based on a geometric perspective of the DB complexity between these two points. Specifically, we propose a novel geodesic-based kernel similarity metric, which we call the Weighted Exponential Geodesic (WEG) kernel. The WEG kernel encodes our expectation that two samples close in Euclidean space may not actually be similar if the DB within a local neighborhood of the samples is highly complex.

Using our similarity formulation, we propose the **G**aussian **P**rocess **E**xplanation Un**C**ertainty (GPEC) framework, which is an instance-wise, model-agnostic, and explainer-agnostic method to quantify the uncertainty of explanations. The proposed notion of uncertainty is complementary to existing quantification methods. Existing methods primarily estimate the uncertainty related to the choice in model parameters and fitting the explainer, which we call *function approximation uncertainty*, and does not capture uncertainty related to the DB. GPEC can combine the DB-based uncertainty with function approximation uncertainty derived from any local feature attribution method.

In summary, we make the following contributions:

- We introduce a geometric perspective on capturing explanation uncertainty and define a novel geodesic-based similarity between explanations. We prove theoretically that the proposed similarity captures the complexity of the decision boundary from a given black-box classifier.
- We propose a novel Gaussian Process-based framework that combines A) uncertainty from decision boundary complexity and B) explainer-specific uncertainty to generate uncertainty estimates for any given feature attribution method and black box model.
- Empirical results show GPEC uncertainty improves understanding of feature attribution methods.

2 Related Works

Explanation Methods. A wide variety of methods have been proposed for the purpose of improving transparency for pre-trained black-box prediction models (Guidotti et al., 2018; Barredo Arrieta et al., 2020). Within this category of *post-hoc* methods, many methods focus on *local* explanations, that is, explaining individual predictions rather than the entire model. Some of these methods generate explanations through local *feature selection* (Chen et al., 2018; Masoomi et al., 2020). In this

work, we focus primarily on *feature attribution* methods, which return a real-valued score for each feature in the sample. For example, LIME (Ribeiro et al., 2016) trains a local linear regression model to approximate the black-box model. Lundberg & Lee (2017) generalizes LIME and five other feature attribution methods using the SHAP framework, which fulfill a number of desirable axioms. While some methods such as LIME and its SHAP variant, KernelSHAP, are model-agnostic, others are designed for specific model architectures, such as neural networks (Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Erion et al., 2021), tree ensembles (Lundberg et al., 2020), or Bayesian neural networks (Bykov et al., 2020), taking advantage of those specific architectures. Another class of method involves the training of separate surrogate models to explain the black-box model (Dabkowski & Gal, 2017; Chen et al., 2018; Schwab & Karlen, 2019; Guo et al., 2018; Jethani et al., 2022). Once trained, surrogate-based methods are typically very fast; explanation generation generally only requires a single inference step from the surrogate model.

Reliability of Explanations. While feature attribution methods have gained wide popularity, a number of issues relating to the reliability of such methods have been uncovered. Alvarez-Melis & Jaakkola (2018) investigate the notion of robustness and show that many feature attribution methods are sensitive to small changes in input. This has been further investigated in the adversarial setting for perturbation-based methods (Slack et al., 2020) and neural network-based methods. (Ghorbani et al., 2019). Kindermans et al. (2019) show that many feature attribution methods are affected by distribution transformations such as those common in preprocessing. The generated explanations can also be very sensitive to hyperparameter choice Bansal et al. (2020). A number of metrics have been proposed for evaluating explainer reliability, such as with respect to adversarial attack (Hsieh et al., 2021), local perturbations (Alvarez-Melis & Jaakkola, 2018; Visani et al., 2022), black-box smoothness (Khan et al., 2022), fidelity to the black-box model (Yeh et al., 2019), or combinations of these metrics (Bhatt et al., 2020).

Explanation Uncertainty. One option for improving the trustworthiness of explainers is to quantify the associated explanation uncertainty. Bootstrap resampling techniques have been proposed as a way to get uncertainty estimates from surrogate-based explainers (Schwab & Karlen, 2019; Schulz et al., 2022). Guo et al. (2018) also proposes a surrogate explainer parameterized with a Bayesian mixture model. Alternatively, Bykov et al. (2020) introduces a method for explaining Bayesian neural networks, which can be transferred to non-Bayesian neural networks. Covert & Lee (2021) derives an unbiased version of KernelSHAP and investigates an efficient way of estimating its uncertainty. Several methods investigate LIME and KernelSHAP in a Bayesian context; for example calculating a posterior over attributions (Slack et al., 2021), investigating the use of priors for explanations (Zhao et al., 2021), or using active learning to improve sampling (Saini & Prasad, 2022).

However, all existing methods for quantifying explanation uncertainty only consider the uncertainty of the explainer as a function approximator. This work introduces an additional notion of uncertainty for explainers that takes into account the uncertainty of the explainer due to the DB.

3 Uncertainty Quantification for Black-Box Explainers

We now outline the GPEC framework, which we parameterize with a Gaussian Process (GP) regression model. We define a vector-valued GP which is trained on data samples as input and explanations as labels. More concretely, consider a data sample $x \in \mathcal{X} \subset \mathbb{R}$ that we want to explain in the context of a black-box prediction model $F : \mathcal{X} \to [0, 1]$, in which the output corresponds to the probability for the positive class. For simplicity, we consider the binary classifier case; we extend to the multi-class case in Section C.3. We apply any given feature attribution method $E : \mathcal{X} \to \mathbb{R}^s$, where s is the dimension of the explanations $e = E(x)$.

Let us draw samples $x_1, \ldots, x_M$ from the data distribution and generate their respective explanations $e_1, \ldots, e_M$. We assume that each explanation e_m is generated from an unobserved latent function $\mathcal{E}$ plus some independent Gaussian noise η_m.

$$e_m = \mathcal{E}(x_m) + \eta_m \quad s.t. \qquad \underbrace{\mathcal{E}(x_m) \sim \mathcal{GP}(0, k(x, x'))}_{\text{Decision Boundary-Aware Uncertainty}} \qquad \underbrace{\eta_m \sim \mathcal{N}(0, \tau_m^{-1})}_{\text{Function Approximation Uncertainty}} \tag{1}$$

where $k(\cdot, \cdot)$ is the specified kernel function and $\tau_m > 0$. We decompose each explanation into two components, $\mathcal{E}(x_m)$ and η_m, which represent two separate sources of uncertainty: 1) a *decision boundary-aware* uncertainty which we capture using the kernel similarity, and 2) a *function approx-*

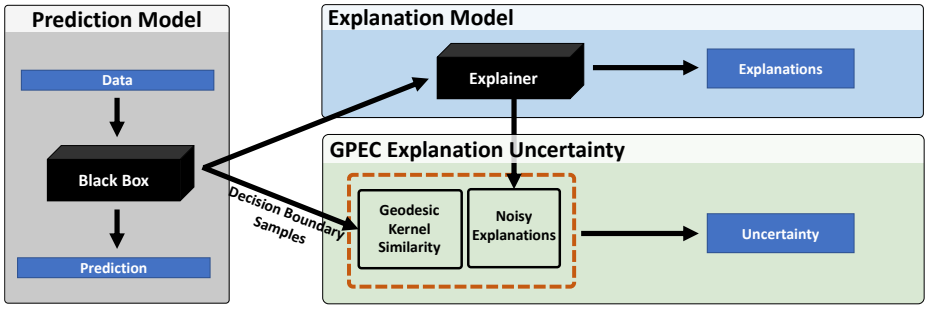

Figure 2: Schematic for the GPEC uncertainty estimation process. GPEC can be used in conjunction with a black-box classifier and explainer to derive an estimate of explanation uncertainty. GPEC takes samples from classifier's decision boundary plus (possibly noisy) explanations from the explainer and fits a Gaussian Progress Regression model with Weighted Exponential Geodesic Kernel. The variance of the predictive distribution combines uncertainty from the black-box classifier complexity and the approximation uncertainty from explainer.

imation uncertainty from the explainer. After specifying $\mathcal{E}(x_m)$ and η_m, we can combine the two sources by calculating the posterior predictive distribution for the GP model; we take the 95% confidence interval of this distribution to be the GPEC uncertainty estimate. Due to space constraints, we provide an overview of GP regression in App. B.

Function Approximation Uncertainty. The η_m component of Eq. 1 represents the uncertainty stemming from explainer specification. For example, this uncertainty can be captured from increased variance from undersampling (perturbation-based explainers) or from the increased bootstrap resampling variance from model misspecification (surrogate-based explainers).

We can therefore define τ based on the chosen explainer. Explainers that include some estimate of uncertainty (e.g. BayesLIME, BayesSHAP, CXPlain) can be directly used to estimate τ. For example, CXPlain returns a distribution of feature importance values for a sample x_m; the variance of this distribution can be directly used as the variance of η_m. For other stochastic explanation methods that do not explicitly estimate uncertainty, we can estimate τ empirically by resampling explanations for the same data sample:

$$\hat{\tau}_m = [\frac{1}{|K|}\sum_{i=1}^{K}(E_i(x_m) - \bar{E}(x_m))^2]^{-1} \quad s.t. \quad \bar{E}(x_m) = \frac{1}{|K|}\sum_{i=1}^{K} E_i(x_m) \tag{2}$$

where each $E_i(x_m)$ is a sampled explanation for x_m. Alternatively, for deterministic explanation methods we can omit the noise and assume that the generated explanations are noiseless.

Decision Boundary-Aware Uncertainty. In contrast, the $\mathcal{E}(x_m)$ component of Eq. 1 draws possible functions from the GP prior that could have generated the observed explanations, with the function space defined by the choice of kernel. Intuitively, the kernel encodes our *a priori* definition of how similar two explanations should be given the similarity of the inputs. In other words, how much information do we expect a given point x to provide for a nearby point x' with respect to their explanations? We use this kernel specification to define a source of uncertainty dependent on the behavior of nearby explanations. In particular, we consider a novel kernel formulation that reflects the complexity of the DB in a local neighborhood of the samples; this is detailed in Section 4. For any given kernel, we can interpret the distribution of possible functions $\mathcal{E}(x_m)$ as an estimate of uncertainty based on the set of previously observed or sampled explanations.

4 Weighted Exponential Geodesic Kernel

Intuitively, the GP kernel encodes the assumption that each individual explanation gives some information about the explanations around it; the choice of kernel defines the neighborhood and magnitude of this shared information. In the GP framework introduced in Section 3, the kernels define the relationship or similarity between two explanations $\mathcal{E}(x)$ and $\mathcal{E}(x')$ based soley on some function of their inputs x and x'. For example, stationary kernels assign the same similarity for any two inputs regardless of where in the data manifold they are located. Instead, we want to encode the assumption

that when the DB is complex, knowing an explanation $\mathcal{E}(x)$ gives limited information about other nearby explanations (see Fig. 3).

4.1 Geometry and Geodesics of the Decision Boundary

We want to relate kernel similarity to the behavior of the black-box model; specifically, the complexity or smoothness of the DB. Given any two points on the DB, the relative complexity of the boundary segment between them can be approximated by the segment length. The simplest form that the DB can take is a linear boundary connecting the two points; this is exactly the minimum distance between the points. As the complexity of the DB grows, there is a general corresponding increase in segment length.

Concretely, we define the DB as a $(d-1)$-dimensional Riemannian manifold embedded in $\mathbb{R}^d$. Given the binary classifier F, we define $\mathcal{M} = \{x \in \mathbb{R}^d : F(x) = \frac{1}{2}\}$ representing the DB[1] of F. Given this interpretation, we can define distances along the DB as geodesic distances in $\mathcal{M}$.

$$d_{geo}(m, m') = \min_{\gamma} \int_0^1 ||\gamma(t)||\mathrm{d}t \quad \forall m, m' \in \mathcal{M} \tag{3}$$

The mapping $\gamma : [0, 1] \to \mathcal{M}$, defined such that $\gamma(0) = m$ and $\gamma(1) = m'$, is a parametric representation of a 1-dimensional curve on $\mathcal{M}$. We can adapt geodesic distance in our kernel selection through the exponential geodesic (EG) kernel Feragen et al. (2015), which is a generalization of the Radial Basis Function (RBF) kernel substituting ℓ_2 distance with geodesic distance:

$$k_{EG}(x, x') = \exp(-\lambda d_{geo}(x, x')) \qquad k_{RBF}(x, x') = \exp(-\lambda ||x - x'||_2^2) \tag{4}$$

The EG kernel has been previously investigated in the context of Riemannian manifolds (Feragen et al., 2015; Feragen & Hauberg, 2016). In particular, while prior work shows that the EG kernel fails to be positive definite for all values of λ in non-Euclidean space, there exists large intervals of $\lambda > 0$ for the EG kernel to be positive definite. Appropriate values can be selected through grid search and cross validation; we assume that a valid value of λ has been selected.

Note that the manifold we are interested in is the DB; applying the EG kernel on the data manifold (i.e. directly using k_{EG} in the GPEC formulation) would not capture model complexity. In addition, naïvely using the EG kernel on the DB manifold would not capture the local complexity with respect to a given explanation; the similarity would be invariant to observed explanations. We therefore need to relate the geodesic distances on $\mathcal{M}$ to samples in the data space $\mathcal{X}$.

4.2 Weighting the Decision Boundary

Consider a probability distribution $p(M)$ with its support defined over $\mathcal{M}$. We weight $p(M)$ according to the ℓ_2 distance between M and a fixed data sample x:

$$q(M|x, \rho) \propto \exp[-\rho||x - M||_2^2]p(M) \tag{5}$$

We evaluate the kernel function $k(x, x')$ by taking the expected value over the weighted distribution.

$$k_{WEG}(x, x') = \int\int \exp[-\lambda d_{geo}(m, m')]\; q(m|x, \rho)\; q(m'|x', \rho)\; \mathrm{d}m\mathrm{d}m' \tag{6}$$

Our formulation is an example of a marginalized kernel (Tsuda et al., 2002): a kernel defined on two observed samples x, x' and taking the expected value over some hidden variable. Given that the underlying EG kernel is positive definite, it follows that the WEG kernel forms a valid kernel.

To evaluate the WEG kernel, we theoretically investigate two properties of the kernel. Theorem 1 shows that the WEG kernel is an extension of the EG kernel for data samples not directly on the DB.

Theorem 1. *Given two points $x, x' \in \mathcal{M}$, then* $\lim_{\lambda\to\infty} k_{WEG}(x, x') = k_{EG}(x, x')$

Proof details are available in App. C.1. Intuitively, as we increase λ, the manifold distribution closest to the points x, x' becomes weighted increasingly heavily. At the limit, the weighting becomes concentrated entirely on the points x, x' themselves, which recovers the EG kernel. Therefore we see that the WEG kernel is simply a weighting of the EG kernel, which is controlled by λ.

[1] Without loss of generality, we assume that the decision rule for the classifier is set to be $\frac{1}{2}$

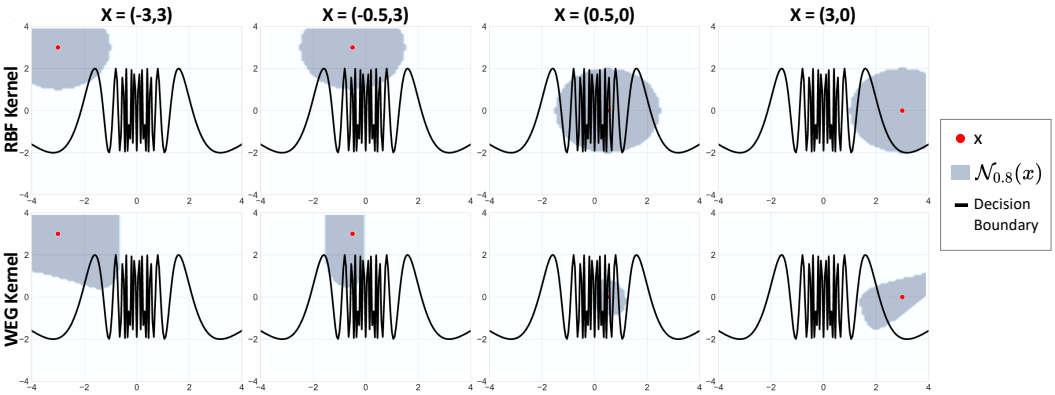

Figure 3: Comparison of RBF and WEG kernel similarity neighborhoods. The gray highlighted region $\mathcal{N}_{0.8}(x) = \{x' : k(x, x') \geq 0.8\}$ indicate points where the kernel similarity is greater than 0.8 with respect to the red point x. The black line is the decision boundary for the classifier $f(x_1, x_2) = 2\cos(\frac{10}{x_1}) - x_2$, where $f(x_1, x_2) \geq 0$ indicates class 1, and $f(x_1, x_2) < 0$ indicates class 0. We see that for GPEC, $\mathcal{N}_{0.8}(x)$ shrinks as x moves closer to segments of the decision boundary which are more complex.

Theorem 2 establishes the negative relationship between WEG kernel similarity and DB complexity. Note that the notation $\text{int}(S)$ indicates the interior of a set S, $\text{f}|_{\mathcal{S}}$ represents the function f restricted to S, and id indicates the identity mapping.

Definition 1. Let $\{U_\alpha\}_{\alpha \in I}$ be charts of an atlas for a manifold $\mathcal{P} \subset \mathbb{R}^d$, where I is a set of indices. Let $\mathcal{P}$ and $\widetilde{\mathcal{P}}$ be differentiable manifolds embedded in $\mathbb{R}^d$, where $\mathcal{P}$ is a Piecewise Linear (PL) manifold. Let $R : \mathcal{P} \to \widetilde{\mathcal{P}}$ be a diffeomorphism. We say $\widetilde{\mathcal{P}}$ is a perturbation of $\mathcal{P}$ on the ith chart if R satisfies the following conditions. There exists a compact subset $K_i \subset U_i$ s.t. $R|_{\mathcal{P} \setminus \text{int}(K_i)} = \text{id}|_{\mathcal{P} \setminus \text{int}(K_i)}$ and $R|_{\text{int}(K_i)} \neq \text{id}|_{\text{int}(K_i)}$. Furthermore there exists a linear homeomorphism between an open subset $\widetilde{U_i} \subseteq U_i$ with $\mathbb{R}^{d-1}$ which contains K_i.

Theorem 2. *Let $\mathcal{P}$ be a $(d$–$1)$-dimension PL manifold embedded in $\mathbb{R}^d$. Let $\widetilde{\mathcal{P}}$ be a perturbation of $\mathcal{P}$ and define $\tilde{k}(x, x')$ and $k(x, x')$ as the WEG kernel defined on $\widetilde{\mathcal{P}}$ and $\mathcal{P}$ respectively. Then $\tilde{k}(x, x') < k(x, x')$ $\forall x, x' \in \mathbb{R}^d$.*

Proof details are in App. C.2. Theorem 2 implies that the linearization of the DB maximizes the kernel similarity; any deviation from the linearization will correspondingly decrease kernel similarity.

To improve the interpretation of the WEG kernel, we apply a normalization to scale the similarity values to be between $[0, 1]$. We construct the normalized kernel k^*_{WEG} as follows:

$$k^*_{WEG}(x, x') = \frac{k_{WEG}(x, x')}{\sqrt{k_{WEG}(x, x) k_{WEG}(x', x')}} \tag{7}$$

4.3 WEG Kernel Approximation

In practice, the integral in Eq. 6 is intractable; we approximate the expected value using Monte Carlo (MC) sampling with K samples:

$$k_{WEG}(x, x') \approx \frac{1}{Z_m Z_{m'} K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \exp[-\lambda d_{geo}(m_i, m_j)] \exp[-\rho(||x - m_i||_2^2 + ||x' - m_j||_2^2)] \tag{8}$$

Z_m, $Z_{m'}$ are the normalization constants for $q(M|x, \rho)$ and $q(M'|y, \rho)$, respectively. We can similarly estimate these values using MC sampling:

$$Z_m = \int \exp[-\rho||x - m||_2^2] p(m) \, \text{d}m \approx \frac{1}{K} \sum_{i=1}^{K} \exp[-\rho||x - m_i||_2^2] \tag{9}$$

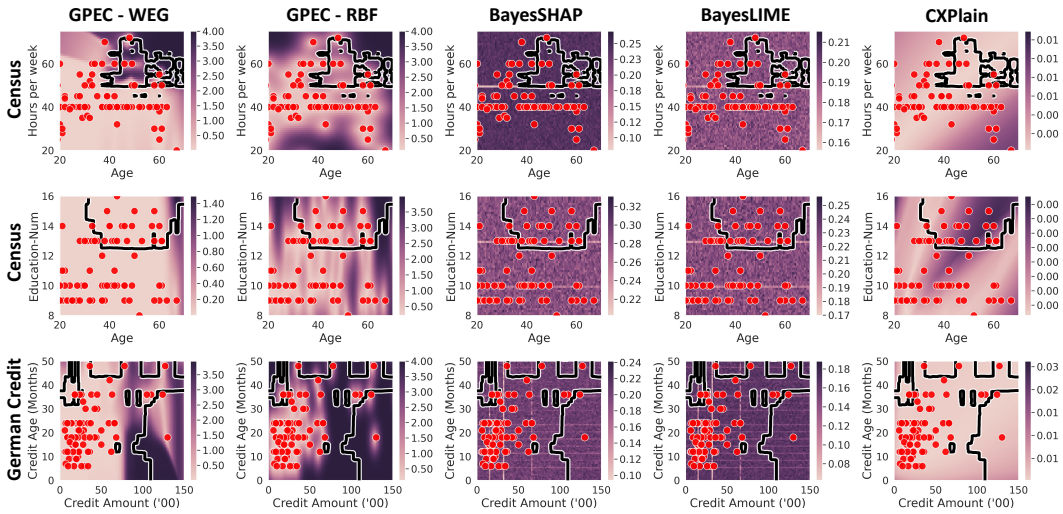

Figure 4: Visualization of estimated uncertainty of explanations for different models and competing methods. The heatmap represents level of uncertainty for a grid of explanations for the feature on the x-axis; dark areas represent high uncertainty whereas lighter areas represent low uncertainty. Red points represent training samples for GPEC and CXPlain, and the reference samples for BayesSHAP and BayesLIME. The black line represents the decision boundary for the XGBoost binary classifier.

GPEC Implementation Details. To use the GPEC framework, we first train the GP regression model with a set of training samples. Once the GP is trained, we can derive uncertainty estimates for test samples from the 95% confidence interval of the posterior predictive distribution. The WEG kernel requires additional setup time to use in the GP. We first calculate the EG kernel matrix by sampling the black-box DB, then weigh each element of the matrix per eq. 5. The estimation of a classifier's DB is an ongoing area of research; a number of methods have been proposed in the literature (e.g. Yan & Xu (2008); Karimi et al. (2019)). Once GPEC is trained, training cost is amortized during inference; estimating uncertainty for test samples using a GP generally has time complexity of $\mathcal{O}(n^3)$, which can be reduced to $\mathcal{O}(n^2)$ using BBMM (Gardner et al., 2018), and further with variational methods (e.g., Hensman et al. (2015)).

5 Experiments

We evaluate the uncertainty estimates from GPEC on a variety of datasets and classifiers. In section 5.2 we compare the DB-aware uncertainty and approximation uncertainty components of GPEC. Section 5.3 evaluates how GPEC captures DB complexity. Section 5.4 evaluates how well GPEC combines sources of uncertainty. Due to space constraints we have additional results in the appendix, including sensitivity analysis for GPEC parameters (F.4), execution time comparison F.1, and additional experiments on combining uncertainty (F.2). All experiments were run on an internal cluster using AMD EPYC 7302 16-Core processors, and all source code will be made public.

5.1 Experimental Setup

Datasets and Models. Experiments are performed on three tabular datasets (Census, Online Shoppers (Sakar et al., 2019), German Credit) from the UCI data repository (Dua & Graff, 2017) and two image datasets (MNIST (LeCun & Cortes, 2010) and fashion-MNIST (f-MNIST) (Xiao et al., 2017)). GPEC can be used with any black-box model; for our experiments we use XGBoost (Chen & Guestrin, 2016) with log-loss for the tabular datasets and a 4-layer Multi-Layered Perception (MLP) model for the image datasets. Additional dataset details are outlined in App. E.1.

Implementation Details. We consider two versions of GPEC: GPEC-WEG which is fit with the WEG kernel, and GPEC-RBF which is fit with the RBF Kernel. Unless otherwise stated, we choose $\lambda = 1.0$ and $\rho = 0.1$ (see App. F.4 for experiments on parameter sensitivity). For the Uncertainty Visualization (Sec. 5.2) and Regularization Test (Sec. 5.3) we use GPEC with the KernelSHAP explainer. The DBPS binary search algorithm (Yan & Xu, 2008) is used to sample from the decision

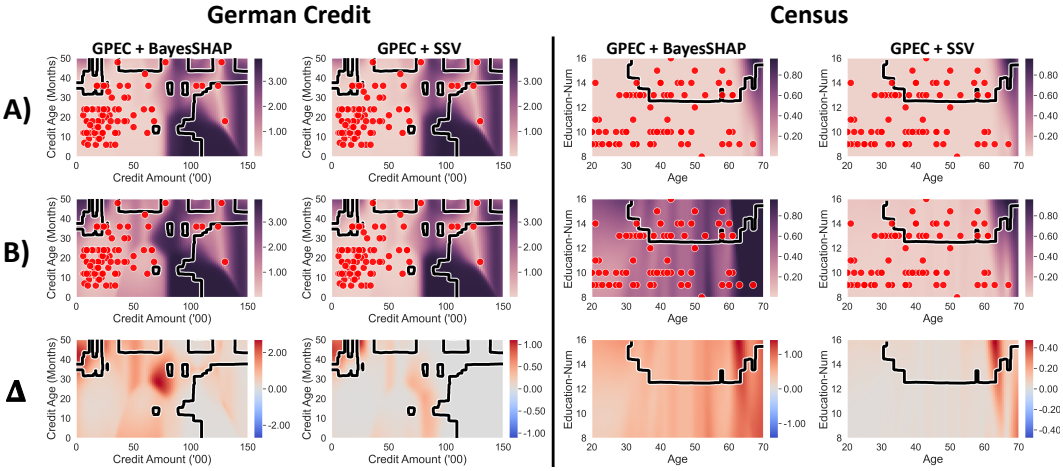

Figure 5: Comparison of the change in uncertainty estimate of explanations when combining sources of uncertainty with GPEC. Row (A) visualizes the estimate using only the WEG Kernel, which captures the DB-based uncertainty assuming noiseless explanations. Row (B) combines the estimate in (A) with an estimate of approximation uncertainty from the given explainer, BayesSHAP and SSV. The third row (Δ) represents the change in uncertainty estimate between (A) and (B).

boundaries of the black-box models. For tabular datasets we apply DBPS to a grid of samples. For image datasets we first apply adversarial attacks to reduce search space size (Karimi et al., 2019). We then apply ISOMAP Tenenbaum et al. (2000), which uses a graph-based approach to estimating geodesic distances. Additional implementation details are available in App. D.

Competing Methods. We compare GPEC to three competing explanation uncertainty estimation methods. BayesSHAP and BayesLIME (Slack et al., 2021) are extensions of KernelSHAP and LIME, respectively, that fit Bayesian linear regression models to perturbed data samples. After fitting, 95% credible intervals can be estimated by sampling the posterior distribution of the feature attributions. CXPlain (Schwab & Karlen, 2019) trains a separate explanation model using a causal-based objective, and applies a bootstrap resampling approach to estimate explanation uncertainty. We report the 95% confidence interval from the set of bootstrapped explanations. Unless otherwise stated, we use the default settings in the provided implementation for all three of these methods.

5.2 Uncertainty Visualization

In order to visualize explanation uncertainty, we train XGBoost binary classifiers using two selected features, which we take to be the black-box model. Both GPEC and competing methods are used to quantify the uncertainty for the feature attributions of the test samples, which we choose to be a grid of 10,000 samples over the data domain. Uncertainty estimates for the x-axis feature are plotted in Figure 4 (results for the y-axis variable are included in App. F.3). Test points are plotted using a heatmap, where darker values indicate high uncertainty. GPEC-WEG, GPEC-RBF, and CXPlain require training samples for their amortized uncertainty estimates; these samples, also used as the reference distribution for BayesSHAP and BayesLIME, are plotted in red. The DB is plotted in black for reference. We observe that the uncertainty heatmap for GPEC-WEG attributes higher uncertainty to samples near nonlinearities in the DB. In contrast, GPEC-RBF provides uncertainty estimates that relate only to the training samples; test sample uncertainty is proportional to distance from the training samples. The methods BayesSHAP, BayesLIME, and CXPlain results in relatively uniform uncertainty estimates over the test samples. CXPlain shows areas of higher uncertainty for Census, however the magnitude of these estimates are small. The uncertainty estimates produced by these competing methods are unable to capture the properties of the black-box model.

5.3 Regularization Test

In this experiment we compare the average uncertainty of explanations as the model is increasingly regularized in order to assess the impact of restricting model complexity. For XGBoost models, we vary the parameter γ, which penalizes the number of leaves in the regression tree functions (Eq. 2 in

Dataset	**Census**			**Online Shoppers**			**German Credit**		
Regularization	γ			γ			γ		
Magnitude	0	5	10	0	5	10	0	5	10
GPEC-WEG	1.573	1.177	1.158	0.209	0.123	0.092	2.665	1.747	0.250
GPEC-RBF	0.493	0.472	0.466	1.740	1.731	1.734	0.153	0.081	0.073
BayesSHAP	0.037	0.037	0.037	0.031	0.031	0.031	0.019	0.019	0.018
BayesLIME	0.097	0.096	0.095	0.098	0.097	0.093	0.085	0.066	0.045
CXPlain	0.064	0.064	0.069	0.004	0.003	0.006	$1.7e^{-4}$	$1.1e^{-4}$	$4.3e^{-4}$

Dataset	**MNIST**						**Fashion MNIST**					
Regularization	ℓ_2			Softplus β			ℓ_2			Softplus β		
Magnitude	0	$1e^{-5}$	$10e^{-5}$	1.0	0.5	0.25	0	$1e^{-5}$	$10e^{-5}$	1.0	0.5	0.25
GPEC-WEG	0.236	0.157	0.078	0.087	0.073	0.056	0.378	0.187	0.063	0.112	0.075	0.061
GPEC-RBF	0.439	0.301	0.232	0.226	0.232	0.236	1.01	0.424	0.247	0.302	0.261	0.250
BayesSHAP	0.025	0.016	0.008	0.013	0.011	0.010	0.030	0.014	0.007	0.018	0.010	0.009
BayesLIME	2.452	1.573	0.737	0.868	0.866	0.721	2.605	1.364	0.666	1.178	0.861	0.779
CXPlain	$0.1e^{-5}$	$5.0e^{-5}$	$8.6\ e^{-5}$	$5.3e^{-5}$	$8.0e^{-5}$	$5.4e^{-5}$	$7.2e^{-5}$	$4.8e^{-5}$	$6.2\ e^{-5}$	$9.0e^{-5}$	$6.6e^{-5}$	$9.6e^{-5}$

Table 1: Average explanation uncertainty estimates over all features for a given classifier with varying levels of regularization. Higher regularization (increasing left to right) generally results in smoother and simpler classifiers.

Chen & Guestrin (2016)). For neural networks, we use two types of regularization. First, we add an ℓ_2 penalty to the weights; this penalty increases as the parameter λ increases. Second, we change the ReLU activation functions to Softplus; a smooth approximation of ReLU with smoothness inversely proportional to a parameter β, which we vary (Dombrowski et al., 2019). We observe in Table 1 that the average uncertainty estimate generated by GPEC-WEG decreases as regularization increases, suggesting that the uncertainty estimates reflect the overall complexity of the underlying black-box model. For the tabular datasets, the estimates for BayesSHAP, BayesLIME, and CXPlain stay relatively flat. Interestingly, the estimates from these methods decrease for the image datasets. We hypothesize that the regularization for the neural network model also increases overall stability of the explanations. GPEC can capture both the uncertainty from WEG kernel and also the estimated uncertainty from the function approximation for the explainer, which we demonstrate in Section 5.4.

5.4 Combining Approximation Uncertainty

Using the WEG kernel with noisy explanation labels enables GPEC to incorporate uncertainty estimates from the explainer. Explainers such as BayesSHAP provide an estimate of uncertainty for their explanations, which can be used as fixed noise in the Gaussian likelihood for GPEC. For explainers with no such estimate, we can empirically resample explanations for the same test sample and take the explanation variance as an estimate of uncertainty. Here, we compare GPEC uncertainty results before and after applying explanation noise estimates. We combine GPEC with two different explainers: BayesSHAP and Shapley Sampling Values (SSV) (Strumbelj & Kononenko, 2013). The two selected explainers are different methods of approximating SHAP values, however the former has an inherent method for quantifying its uncertainty while the latter does not. In Figure 5 row (A) we plot the heatmap for uncertainty estimates using GPEC using only the WEG Kernel and no added noise from the explainers. In row (B) we add the additional uncertainty estimates from BayesSHAP and SSV. The difference in quantified uncertainty is visualized in the third row (Δ). We observe that combining the uncertainty estimates from BayesSHAP and SSV in the GPEC Gaussian Process formulation increases the overall uncertainty estimate.

6 Limitations and Conclusion

Generating uncertainty estimates for feature attribution explanations are essential for building reliable explanations. We introduce a novel GP-based approach that can be used with any black-box classifier and feature attribution method to generate uncertainty estimates that capture the complexity of the black-box model. Experiments show that capturing this type of uncertainty improves understanding of the explanations and the black-box model itself. Regarding limitations, GPEC relies on the ability to approximate the black-box DB, which increases the training time. The effect of this is minimized during inference, as the DB only needs to be sampled during training. Additionally, in its current implementation GPEC is limited to black-box classifiers; we leave the extension to regression models as future work.

REFERENCES

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

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, 2015.

Naman Bansal, Chirag Agarwal, and Anh Nguyen. SAM: The Sensitivity of Attribution Methods to Hyperparameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8673–8683, 2020.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations. *arXiv:2005.00631 [cs, stat]*, May 2020.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society, 2001.

Kirill Bykov, Marina M-C Höhne, Klaus-Robert Müller, Shinichi Nakajima, and Marius Kloft. How much can I trust You?–Quantifying uncertainties in explaining neural networks. *arXiv preprint arXiv:2006.09000*, 2020.

George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0-534-24312-6.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 883–892. PMLR, July 2018.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

Ian Covert and Su-In Lee. Improving kernelSHAP: Practical shapley value estimation using linear regression. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3457–3465. PMLR, April 2021.

Ian Covert, Scott Lundberg, and Su-In Lee. Feature Removal Is a Unifying Principle for Model Explanation Methods. *arXiv:2011.03623 [cs, stat]*, November 2020.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, July 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00343-w.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Aasa Feragen and Søren Hauberg. Open Problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel? In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1647–1650, Columbia University, New York, New York, USA, June 2016. PMLR.

Aasa Feragen, François Lauze, and Søren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3032–3042, 2015. doi: 10.1109/CVPR.2015.7298922.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33013681.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *Acm Computing Surveys*, 51 (5), August 2018. ISSN 0360-0300. doi: 10.1145/3236009.

Wenbo Guo, Sui Huang, Yunzhe Tao, Xinyu Xing, and Lin Lin. Explaining deep learning models – a bayesian non-parametric approach. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/hensman15.html.

Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and Methods for Explanation through Robustness Analysis. *arXiv:2006.00442 [cs, stat]*, April 2021.

Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022.

Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460*, 2019.

Zulqarnain Khan, Aria Masoomi, Davin Hill, and Jennifer Dy. Analyzing the effects of classifier lipschitzness on explainers, 2022.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pp. 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Kivanc Kose, Alican Bozkurt, Christi Alessi-Fox, Melissa Gill, Caterina Longo, Giovanni Pellacani, Jennifer G Dy, Dana H Brooks, and Milind Rajadhyaksha. Segmentation of cellular patterns in confocal images of melanocytic lesions in vivo via a multiscale encoder-decoder network (med-net). *Medical Image Analysis*, 67:101841, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13374–13386. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9b10a919ddeb07e103dc05ff523afe38-Paper.pdf.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL https://doi.org/10.7551/mitpress/3206.001.0001.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL http://arxiv.org/abs/1707.04131.

Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. doi: 10.21105/joss.02607. URL https://doi.org/10.21105/joss.02607.

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*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Co., New York, third edition, 1976.

Aditya Saini and Ranjitha Prasad. Select wisely and explain: Active learning and probabilistic local post-hoc explainability. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 599–608, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534191.

C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10):6893–6908, October 2019. ISSN 1433-3058. doi: 10.1007/s00521-018-3523-0.

Jonas Schulz, Raul Santos-Rodriguez, and Rafael Poyiadzi. Uncertainty quantification of surrogate explanations: An ordinal consensus approach. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 3, 2022.

Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 490–494, 2020. doi: 10.1109/ICESC48915.2020.9155614.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, August 2017.

Ashutosh Singh, Christiana Westlin, Hedwig Eisenbarth, Elizabeth A. Reynolds Losin, Jessica R. Andrews-Hanna, Tor D. Wager, Ajay B. Satpute, Lisa Feldman Barrett, Dana H. Brooks, and Deniz Erdogmus. Variation is the norm: Brain state dynamics evoked by emotional video clips. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 6003–6007, 2021. doi: 10.1109/EMBC46164.2021.9630852.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375830.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable Post hoc Explanations Modeling Uncertainty in Explainability. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–2323, December 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2319.

Koji Tsuda, Taishin Kin, and Kiyoshi Asai. Marginalized kernels for biological sequences. *Bioinformatics (Oxford, England)*, 18(suppl$_1$):S268–S275, July 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.suppl_1.S268.

Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022. doi: 10.1080/01605682.2020.1865846.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Zhiyong Yan and Congfu Xu. Using decision boundary to analyze classifiers. In *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, volume 1, pp. 302–307, Nov 2008. doi: 10.1109/ISKE.2008.4730945.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and Sensitivity for Explanations. *arXiv:1901.09392 [cs, stat]*, November 2019.

Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. BayLIME: Bayesian local interpretable model-agnostic explanations. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 887–896. PMLR, July 2021.