But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI

Charlie Marx
Stanford University

Youngsuk Park
AWS AI Labs

Hilaf Hasson
AWS AI Labs

Yuyang Wang
AWS AI Labs

Stefano Ermon
Stanford University

Jun Huan
AWS AI Labs

Abstract

Although black-box models can accurately predict outcomes such as weather patterns, they often lack transparency, making it challenging to extract meaningful insights (such as which atmospheric conditions signal future rainfall). Model explanations attempt to identify the essential features of a model, but these explanations can be inconsistent: two near-optimal models may admit vastly different explanations. In this paper, we propose a solution to this problem by constructing uncertainty sets for explanations of the optimal model(s) in both frequentist and Bayesian settings. Our uncertainty sets are guaranteed to include the explanation of the optimal model with high probability, even though this model is unknown. We demonstrate the effectiveness of our approach in both synthetic and real-world experiments, illustrating how our uncertainty sets can be used to calibrate trust in model explanations.

1 Introduction

Data is now collected at a much faster rate than can be processed directly by humans. Thus, machine learning has been used to synthesize complex datasets into predictive models. For example, models can predict the 3D structure of proteins from their amino acid sequences (Jumper et al., 2021) and forecast supply chain demand (Carbonneau et al., 2008; Sharma et al., 2020). However, modern models are often black-box in nature, meaning that even when they make accurate predictions, it is difficult to extract interpretable principles or intuitions. Whereas human experts can communicate their reasoning, predictive models typically lack the ability to communicate principles.

In response to this challenge, there has been growing interest in model *explanations*: human-interpretable descriptions of model predictions (Koh and Liang, 2017; Ribeiro et al., 2018; Simonyan et al., 2013; Sundararajan et al., 2017). The explanations highlight aspects of the model that are particularly relevant for some downstream goal, such as calibrating trust in a model or identifying patterns in complex data. Popular explanations include SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), integrated gradients (Chattopadhyay et al., 2019), TCAV (Kim et al., 2018), and counterfactual explanations (Ustun et al., 2019).

Use cases for model explanations can be organized around two goals: model auditing and scientific inquiry. In model auditing, the goal is to validate or debug the predictions of a trained model. For example, we might ask *"In what way does this climate model for global surface temperature depend on CO_2 emissions?"* In contrast, in scientific inquiry the object of interest is the data generating distribution itself. An analogous question for scientific inquiry would be *"In what way is the global surface temperature explained by CO_2 emissions?"* Explanations used for model audit give insights *about the model*, whereas explanations for scientific inquiry give insights *about the world*. When the model is suboptimal or there are multiple near-optimal models, the explanations of a model and can be quite different from those of the data generating distribution. Consequently, explaining a single model may reveal little about the process the model is approximating.

Explanations are already being used for scientific inquiry in many domains, such as materials discovery (Raccuglia et al., 2016), genomics (Bi et al., 2020; Johnsen et al., 2021), motor vehicle collisions (Wen et al., 2021), economics (Jabeur et al., 2021; Mokhtari et al., 2019), and environmental science (Zhou et al., 2022). Usually, a practitioner chooses a single "best-fitting" model and treats explanations of that model as representative of the data generating distribution. However, model explanations are known to be unstable (i.e., sensitive to small perturba-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

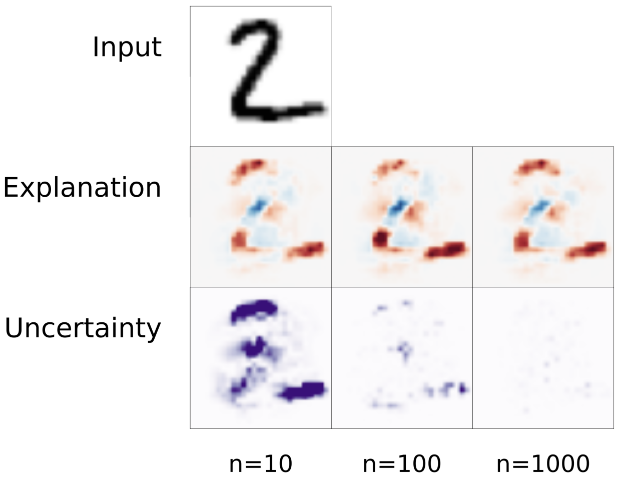

Figure 1: **Top:** This image of a ‘2’ digit is given as input to neural networks trained on $n = 10, 100$, and 1000 MNIST digit examples. Each model predicts the probability the image is of each possible digit, 0 through 9. **Middle:** Feature attribution scores computed using Deep SHAP (Lundberg and Lee, 2017; Shrikumar et al., 2017). The color represents the impact of each pixel on the predicted probability assigned to the correct label ‘2’ (red is positive impact, blue is negative impact). **Bottom:** Uncertainty estimates for the attribution of each pixel, computed using our conformal explanation intervals method. The uncertainty of a pixel's attribution is measured as the difference between the maximum and minimum plausible attribution. Darker colors represent higher uncertainty. Uncertainty decreases as the number of training examples increases.

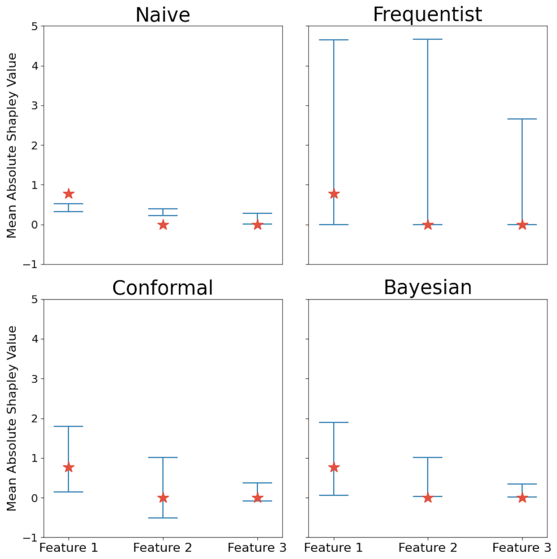

Figure 2: A comparison of methods for computing uncertainty sets for explanations. The explanation is the mean absolute Shapley value, a measure of feature importance. In each panel, the feature importance for the true model that generated the data is marked by a red star. The three features follow a multivariate Gaussian distribution and the first two features are highly correlated. The true labels were sampled from the linear model $y^{(i)} = [1, 0, 0] \cdot x^{(i)} + \epsilon^{(i)}$, where $\epsilon^{(i)} \overset{iid}{\sim} \mathcal{N}(0, 1)$. In the top left panel, we subsample the dataset 100 times and explain a best-fitting linear model for each dataset. Note that the best-fitting model consistently underestimates the importance of Feature 1. In the other three panels, we display confidence intervals generated by the three methods we propose. The frequentist intervals come with strong guarantees, but tend to be wider. The conformal and Bayesian approaches take advantage of additional information (e.g., prior and posterior distribution) to get tighter intervals.

tions in the data) (Adebayo et al., 2018; Alvarez-Melis and Jaakkola, 2018; Dombrowski et al., 2019; Ghorbani et al., 2019; Lakkaraju et al., 2020; Slack et al., 2020) and inconsistent (i.e., random variations in training algorithms can lead models trained on the same data to give different explanations) (Lee et al., 2019). The problem is worsened by the phenomenon of *model multiplicity*: the existence of distinct models with comparable performance (Black et al., 2022; D'Amour et al., 2020; Marx et al., 2020). If there exist competing models—each of which provides a different explanation of the data-generating distribution (Breiman, 2001)—how can we tell which explanation is correct? These issues threaten the applicability of existing explainability procedures for scientific inquiry. Given that explanations are known to vary widely among even near-optimal models (Dong and Rudin, 2019), we cannot assume an explanation from a model with good performance is representative of the data generating distribution.

In this work, we aim to quantify the degree to which an explanation can be used for valid scientific inquiry. We develop broadly applicable wrappers that provide uncertainty estimates for existing explainability methods. Instead of computing the explanation for a single best-fitting model, we want to infer the explanation of the population optimal model. For ease of language, we refer to this explanation as the “optimal explanation”. Here, “optimal” simply means that the explanation is from the optimal model. Since the optimal model is not known, we return an *explanation set* with the guarantee that the optimal explanation is included in the explanation set with high probability (e.g., 95%). See Figure 3 for a high-level illustration of our approach.

When we have a well-specified probabilistic model and we evaluate models using a proper loss function, the data generating distribution is an optimal model. In this setting, our explanation sets can be viewed as uncertainty sets for the explanation from the data generating distribution. In Figure 2, naive uncertainty sets constructed by explaining multiple models consistently disagree with the true expla-

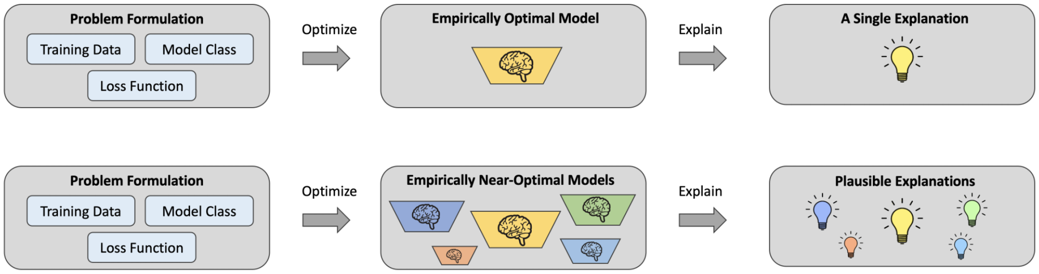

Figure 3: **Top:** The standard explainability pipeline. A single "best-fitting" model is trained and a single explanation is generated from this model. **Bottom:** An instantiation of our proposed explainability pipeline. We explore the set of explanations associated with near-optimal models to construct a confidence set for the explanation of the true model.

nation of a well-specified linear model. In contrast, our uncertainty sets include the true explanation.

Our main contributions include:

- We present a framework for explaining the (unknown) optimal model, as opposed to a trained model. We give a simple example where existing explainability procedures fail to recover the optimal explanation in this framework.
- We propose three simple yet rigorous methods to construct uncertainty sets for optimal explanations: one for a frequentist setting, and two for a Bayesian setting under different assumptions. We provide finite-sample coverage guarantees for the uncertainty sets given by each method.
- Through simulations, we demonstrate the effectiveness of our method in terms of the coverage, i.e., how often the uncertainty set includes the true explanation, and the size of the uncertainty set. We also apply our methods to real datasets to infer feature importance.

The rest of the paper is organized as follows. After reviewing related works (Section 2), we introduce a framework for quantifying uncertainty in model explanations (Section 3). We then develop frequentist and Bayesian approaches to construct principled uncertainty sets (Sections 4 and 5, respectively). Finally, we conduct an experimental study on both synthetic and real-world datasets (Section 6), followed by a final discussion (Section 7).

2 Related Work

Explainable AI. Explainable AI (XAI) aims to present model behavior in a way that humans can easily understand. Some models are inherently more interpretable, such as generalized linear models (GLMs) (Nelder and Wedderburn, 1972) and tree-based models (Sagi and Rokach, 2020). For less interpretable models, post-hoc explanations can still provide insights. Popular methods include Shapley-value based approaches (Frye et al., 2020; Heskes et al., 2020; Lundberg and Lee, 2017; Shapley, 1953), perturbation-based approaches (Fong and Vedaldi, 2019), local approximations (Ribeiro et al., 2016), tree-based methods (Chen and Guestrin, 2016), and DeepLIFT (Shrikumar et al., 2017). Recently, there have been attention to improving the robustness of explanations to distribution shifts (Lakkaraju et al., 2020; Ning et al., 2022). Separately, for probabilistic models, several studies explain uncertainty estimates (Antorán et al., 2020; Ley et al., 2021) and their effects (Shaikhina et al., 2021).

Causal Inference for Explanation. In parallel to XAI, causal inference attempts to understand the world by identifying causal relationships from data. The popular potential outcomes framework (Holland, 1986; Rubin, 1974; Splawa-Neyman et al., 1990) and causal graphical models (Pearl, 1988) typically require either some control of the experiments (like randomized trials), or causal assumptions such as unconfoundedness. These methods can enable scientific discovery, but require more care to be used correctly. In contrast, most XAI methods can be deployed to any accessible predictive models. Recently, several works have considered causal feature relevance (Heskes et al., 2020), and causal contributions (Janzing et al., 2020a), bringing causal inference and explainability closer together (Janzing et al., 2020b). However, these causal explanation methods are still centered around causal interpretations of a trained model rather than inferring the true data generating distribution.

Uncertainty Quantification. There are many ways to quantify uncertainty in prediction tasks, including via uncertainty sets or by providing a probability distribution over potential outcomes. A few popular methods include Gaussian Processes (Barber, 2012; Bishop and Nasrabadi, 2006), which predict full probability distributions, and quantile regression (Koenker, 2005; Koenker and Bassett Jr, 1978; Park et al., 2022), which can give prediction intervals by minimizing the pinball loss. Conformal prediction is a post-hoc process that can construct valid uncertainty sets or predictive distributions from heuristic notions of uncertainty (Angelopoulos and Bates, 2021; Shafer and Vovk, 2008). To the best of our knowledge, conformal prediction has only ever been applied in a frequentist setting.

Uncertainty in Explanations Several existing works consider uncertainty in model explanations. Slack et al. (2021) develop Bayesian methods for quantifying uncertainty about the explanations of a single trained model. Dong and Rudin (2019) give methods to compute the set of plausible variable importances for a restricted class of models. Our work differs from this work in that we aim to construct general purpose methods to quantify uncertainty about the explanations for the data generating distribution.

3 Framework

3.1 Preliminaries

We consider the task of using features $x \in \mathcal{X}$ to predict an outcome $y \in \mathcal{Y}$. Given a dataset of n i.i.d. pairs $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the learning task is to select a probabilistic model f from a model class $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{P}(\mathcal{Y})\}$ that approximates the conditional distribution for y given x. Here, $\mathcal{P}(\mathcal{Y})$ is the set of probability distributions over $\mathcal{Y}$. Given a loss function $\ell(f(x), y)$, our goal is to minimize the expected loss $\mathcal{L}(f) = \mathbb{E}\left[\ell(f(x), y)\right]$.

Let $f^* \in \mathcal{F}$ be a model that minimizes the expected loss, so $\mathcal{L}(f^*) \leq \mathcal{L}(f)$ for all $f \in \mathcal{F}$. We assume the model is well-specified, so there exists some model $f \in \mathcal{F}$ that outputs the true conditional distribution $p(y \mid x)$. Furthermore, we assume the loss is a strictly proper scoring rule, so predicting the true conditional distribution is optimal. Under these assumptions, we refer to f^* as the *true model* since it exactly reflects the data generating distribution. When our model is misspecified, f^* is instead the optimal model from within our model class.

Since we do not know the true model, we use some model-fitting algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$, where $\mathcal{D} = (\mathcal{X} \times \mathcal{Y})^n$, that takes as input a dataset D and outputs a model $\hat{f} = \mathcal{A}(D)$. For example, in empirical risk minimization we choose the model $\hat{f}$ that minimizes the loss on the training data.

3.2 Model Explanations

We are interested in an explanation function $\phi : \mathcal{F} \to \Phi$ that assigns to every model an explanation in some space Φ. The explanation $\phi(f)$ can be a simple function of f, such as the predicted conditional mean for a single input x, $\phi_x^{\text{mean}}(f) = \mathbb{E}_{y \sim f(x)}\left[y\right]$, or a more complex function. For example, we can consider $\phi_{i,x}^{\text{shap}}(f)$, the Shapley value of the i-th feature applied to the feature vector x, with D as the reference dataset; or to the average absolute Shapley value of the i-th feature $\phi_i^{\text{shap}}(f) := \mathbb{E}_{x \sim D}[|\phi_{i,x}^{\text{shap}}(f)|]$.

In binary classification where $y \in \{0, 1\}$, one can consider a counterfactual explanation $\phi_{x,+}^{\text{CF}}(\hat{f})$ that returns the closest point x' to x such that the label is predicted to be most likely of the positive class $\hat{P}(Y = 1 \mid X = x') > 0.5$.

We are interested in the explanation of the true model $\phi(f^*)$. When ϕ is a simple explanation such as the conditional mean ϕ_x^{mean}, we may be able to directly estimate $\phi(f^*)$ using standard statistical techniques. When $\phi(f^*)$ is difficult to estimate directly (e.g., the Shapley values of the true model), we can first estimate the true model f^* then apply the explanation ϕ.

3.3 Quantifiying Uncertainty for Explanations

However, the explanation of our trained model $\phi(\hat{f})$ could be meaningfully different than the true explanation $\phi(f^*)$. For example, consider the conditional mean explanation $\phi_x^{\text{mean}}(\hat{f})$ for some rare input x taken from our dataset. An expressive model class could vary $\hat{f}(x)$ drastically without changing any other predictions on the dataset (and therefore only minimally change the loss). Thus, it is not enough to simply report $\phi(\hat{f})$; we instead need to quantify our uncertainty about $\phi(f^*)$.

In this work we produce *uncertainty sets* for the explanation of the true model. Using the data D, we construct an uncertainty set $C = C(D)$ that is guaranteed to include the true explanation with high probability

$$\mathbb{P}\left(\phi(f^*) \in C\right) \geq 1 - \alpha, \tag{1}$$

for some desired confidence level $1 - \alpha$ with $\alpha \in (0, 1)$. In Equation (1), the uncertainty set C is random due to its dependence on the data D. In Section 5, we consider Bayesian models, where the model itself is a random variable. By convention, in the Bayesian perspective we will denote f^* by f instead to indicate that the data generating distribution is random.

From the Bayesian perspective, with an additional assumption that the posterior can be sampled exactly (see Section 5.1), one can achieve the following guarantee by employing credible intervals:

$$\mathbb{P}\left(\phi(f) \in C \mid D\right) \geq 1 - \alpha \tag{2}$$

For when we cannot access exact samples from the posterior (Section 5.2), we propose an algorithm inspired by conformal prediction to recover a weaker guarantee, $\mathbb{P}\left(\phi(f) \in C\right) \geq 1 - \alpha$, where we no longer condition on the data. This weaker coverage guarantee is over the prior rather than the posterior, unlike a typical Bayesian result.

Finally, we compare all three methods in Section 6. In our analysis we focus on two metrics: how often the uncertainty set includes the true explanation (the "coverage"), and the size of the uncertainty set. In general, higher coverage and tighter uncertainty sets are preferable.

Method	Requires Prior	Requires Posterior	Guarantee
Frequentist	No	No	$\mathbb{P}\left(\phi(f^*) \in C\right) \geq 1-\alpha$
Conformal	Yes	No	$\mathbb{P}\left(\phi(f) \in C\right) \geq 1-\alpha$
Bayesian	Yes	Yes	$\mathbb{P}\left(\phi(f) \in C \mid D\right) \geq 1-\alpha$

Table 1: A comparison of the three methods we propose. The frequentist approach assumes there exists a fixed, true model f^* and gives a confidence interval that includes the explanation for the true model with high probability. The randomness in the frequentist guarantee comes exclusively from the dependence of the confidence interval on the data. In the conformal and Bayesian approaches, we treat the model as a random variable f distributed according to some prior distribution. Thus, the randomness in the guarantees for the conformal and Bayesian approaches is over the model f, the data D, and additional simulated randomness in C we use to obtain the exact guarantee.

4 Frequentist Explanation Interval

In this section, we introduce a method for constructing valid confidence intervals in a frequentist setting when the model class is *sufficiently simple*. We measure simplicity in a learning-theoretic sense; our results hold for model classes that satisfy uniform convergence.

Uniform convergence states that the empirical loss $\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$ converges to the population loss $\mathcal{L}(f) = \mathbb{E}\left[\ell(f(x), y)\right]$ "uniformly" across the model class as the number of training samples n goes to infinity.

Definition 1. *A model class $\mathcal{F}$ has the* uniform convergence *property if, for any distributions P over $\mathcal{X} \times \mathcal{Y}$, any error rate $\alpha > 0$, and any tolerance $\epsilon > 0$, there exists a sample size $n < \infty$ such that*

$$\mathbb{P}_{D\sim P}\left(\sup_{f\in\mathcal{F}} |\mathcal{L}(f) - \mathcal{L}_n(f)| \leq \epsilon\right) \geq 1-\alpha. \tag{3}$$

We say that $\mathcal{F}$ satisfies (α, ϵ_n)-uniform convergence if n is a sufficiently large sample size to achieve the inequality in Equation (3) with α and $\epsilon = \epsilon_n$. See examples of well-known uniform convergence results in Appendix B.2. First, we will note that uniform convergence gives us a confidence set for the true model. Then, we will bound the explanation of the true model by computing the most extreme explanations within this confidence set.

An immediate result of uniform convergence is that the true model has bounded excess empirical loss.

Lemma 1. *If $\mathcal{F}$ satisfies (α, ϵ_n)-uniform convergence, then with probability at least $1-\alpha$,*

$$\mathcal{L}_n(f^*) \leq \inf_{f\in\mathcal{F}} \mathcal{L}_n(f) + 2\epsilon_n. \tag{4}$$

See Appendix A for a simple proof of Lemma 1. This bound gives us a confidence set for the true model,

$$\mathcal{F}_\alpha = \left\{f \in \mathcal{F} : \mathcal{L}_n(f) \leq \inf_{f'\in\mathcal{F}} \mathcal{L}_n(f') + 2\epsilon_n\right\}, \tag{5}$$

which includes the true model with probability at least $1-\alpha$. Thus, the set of explanations corresponding to $\mathcal{F}_\alpha$, namely $C_{\text{freq}} = \{\phi(f) : f \in \mathcal{F}_\alpha\}$, satisfies Equation (1).

Proposition 1. *Suppose $\mathcal{F}$ is well-specified and satisfies (α, ϵ_n)-uniform convergence. Then the confidence interval $C_{freq} = \{\phi(f) : f \in \mathcal{F}_\alpha\}$ includes the true explanation with probability at least $1-\alpha$.*

$$\mathbb{P}\left(\phi(f^*) \in C_{freq}\right) \geq 1-\alpha \tag{6}$$

The randomness in Equation (6) is over the dataset used to compute C_{freq}. Proposition 1 provides a very general guarantee that applies to all possible $f^* \in \mathcal{F}$ and any data distribution. This generality typically comes at the cost of larger confidence sets. Computing C_{freq} exactly can be difficult in practice, depending on the model class. In the following Section 4.1, we elaborate on this challenge and propose an algorithm for efficiently approximating C_{freq}.

4.1 Computing Confidence Set via Pareto Frontier

For simplicity, we now consider a real-valued explanation, so $\Phi = \mathbb{R}$. However, the methods described in this section can easily be extended to vector-valued explanations. One option is to construct confidence intervals that hold marginally for each component. Another is to apply a union bound to get confidence sets that hold jointly for the entire vector.

Note that C_{freq} is a subset of the interval $\left[\inf_{f\in\mathcal{F}_\alpha} \phi(f), \sup_{f\in\mathcal{F}_\alpha} \phi(f)\right]$. In fact, if $\mathcal{F}_\alpha$ is connected and ϕ is continuous, then the sets are equal up to measure 0. In turn, estimating the endpoints of this interval amounts to solving non-convex optimization problems, which can be difficult to solve exactly:

$$\text{minimize} \quad \phi(f) \quad \text{s.t.} \quad f \in \mathcal{F}_\alpha \tag{7}$$

$$\text{maximize} \quad \phi(f) \quad \text{s.t.} \quad f \in \mathcal{F}_\alpha \tag{8}$$

However, we can solve a set of related unconstrained problems to approximate the solution. For Equation (7), we can define a mixed training objective:

$$J_\lambda(f) = \lambda\phi(f) + (1-\lambda)\mathcal{L}_n(f) \tag{9}$$

By optimizing this objective for a sequence of $\lambda \in [0, 1]$, we can estimate the Pareto frontier of $\phi(f)$ and $\mathcal{L}_n(f)$ (and of $-\phi(f)$ and $\mathcal{L}_n(f)$, by flipping a sign). By choosing the first point on this Pareto frontier that satisfies the constraint, we can estimate the solution to the optimization problems posed in Equations (7) and (8). While not exact, these solutions provides an upper bound to the solution for Equation (7) and a lower bound for the solution to Equation (8).

```
Algorithm 1: UCEI: Uniform Convergence Explanation Intervals
Input : dataset $D$, mixture weights
        $0 \leq \lambda_1 < \cdots < \lambda_K \leq 1$
Estimate the ERM $\hat{f} = \min_{f \in \mathcal{F}} \mathcal{L}_n(f)$ and its
  empirical risk $\mathcal{L}_n(\hat{f})$
for $\lambda \in \{\lambda_1, \dots, \lambda_K\}$ do
    Optimize the mixed objective
      $\hat{f}_\lambda^- = \arg\min_{f \in \mathcal{F}} \quad \lambda\phi(f) + (1-\lambda)\mathcal{L}_n(f)$
    Optimize the mixed objective
      $\hat{f}_\lambda^+ = \arg\min_{f \in \mathcal{F}} -\lambda\phi(f) + (1-\lambda)\mathcal{L}_n(f)$
end
Return:
The confidence interval $\hat{C}_{\text{freq}}$ with lower bound
    $\min\{\phi(\hat{f}_\lambda^-) : \mathcal{L}_n(f_\lambda^-) \leq \mathcal{L}_n(\hat{f}) + 2\epsilon_n\}$
and upper bound
    $\max\{\phi(\hat{f}_\lambda^+) : \mathcal{L}_n(f_\lambda^+) \leq \mathcal{L}_n(\hat{f}) + 2\epsilon_n\}$
```

The frequentist guarantee relies on our ability to exactly solve the optimization problems in Equations (7) and (8). When ϕ is differentiable, as is the case for popular methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), we can use backpropagation to optimize the mixed training objective. One can also optimize each mixed objective J_λ in parallel, or adaptively search for a λ that gives a model with empirical loss close to $\min_{f \in \mathcal{F}} \mathcal{L}_n(f) + 2\epsilon_n$.

5 Bayesian Explanation Sets

The algorithm in the previous section guarantees coverage for any true function. A natural question to ask is, "can we get tighter uncertainty sets if we instead require coverage *on average*, when the true model is distributed according to some known distribution?" Consider a Bayesian model, where instead of estimating a fixed but unknown true function f^*, we assume the model follows a prior distribution $p(f)$. We are then interested in the posterior distribution $p(f \mid D)$, which represents our updated beliefs about the model after observing the data. A *credible set* for the posterior distribution is any subset of $\mathcal{F}$ that has probability at least $1-\alpha$ under the posterior. We can similarly define a credible set for the explanation $\phi(f)$ as any subset of Φ that includes the explanation of a model drawn from the posterior with probability at least $1-\alpha$.

5.1 Bayesian Models with a Posterior Sampler

First, we consider the case where we have sample access to the posterior distribution, i.e., a sample f_t can be drawn from exactly $p(f \mid D)$ (without approximation). Credible intervals give us a natural notion of uncertainty quantification for explanations of Bayesian models; we want a set of explanations C_{Bayes} that satisfies the following inequality:

$$\mathbb{P}\left(\phi(f) \in C_{\text{Bayes}} \mid D\right) \geq 1-\alpha \tag{10}$$

Below, we describe Algorithm 2, which outputs C_{Bayes} achieving the guarantee in Equation (10). Suppose that we have T models $f_1, \dots, f_T$ sampled independently from the posterior distribution $p(f \mid D)$. We can explain each model to get T explanations $\phi(f_1), \dots, \phi(f_T)$, which are independently distributed according to the posterior for the explanation $p(\phi(f) \mid D)$. We can then use these samples the estimate the quantiles of $p(\phi(f) \mid D)$. The quantiles of $p(\phi(f) \mid D)$ tell us how to construct credible intervals for the explanation. For example, the interval between the 0.05 and 0.95 quantiles of $p(\phi(f) \mid D)$ represents a credible interval with 90% probability under the posterior distribution. We cannot infer the quantiles of $p(\phi(f) \mid D)$ exactly from T samples, but we can estimate the quantiles in such a way as to guarantee Equation (10) holds with a finite number of samples T, and not only asymptotically. To see this, consider drawing one more model from the posterior $f_{T+1} \sim p(f \mid D)$. Then $\phi(f_1), \dots, \phi(f_T), \phi(f_{T+1})$ are i.i.d. explanations. It follows that $\phi(f_{T+1})$ is equally likely to be the smallest, second smallest, ..., largest element of this collection. If we define the ranking function $R(u) = \sum_{t=1}^{T} \mathbb{1}\{u \leq \phi(f_t)\}$ then $R(\phi(f_{T+1}))$ is distributed uniformly on the set $\{0, 1, 2, \dots, T\}$. Thus, if we define the interval C_{Bayes} with lower bound and upper bound as the $\lfloor \frac{\alpha}{2}(T+1) \rfloor / T$-quantile and $\lceil \left(1-\frac{\alpha}{2}\right)(T+1) \rceil / T$-quantile (respectively) of the set $\{\phi(f_1), \dots, \phi(f_T)\}$, then Equation (10) is guaranteed to hold, even in the finite-data regime. This is because C_{Bayes} is random, even conditioned on the data D, since C_{Bayes} also depends on the T randomly drawn models from the posterior.

```
Algorithm 2: BEI: Bayesian Explanation Intervals
Input : Sampler of posterior distribution $p(f \mid D)$,
        explanation algorithm $\phi$, the number of
        samples $T$
for $t = 1, \dots, T$ do
    Sample a model $f_t \sim p(f \mid D)$
    Compute an explanation $\phi(f_t)$ for the sampled
      model
end
Return:
the confidence interval $C_{\text{Bayes}}$ with lower bound
Quantile$(\{\phi(f_1), \dots, \phi(f_T)\}; \lfloor \frac{\alpha}{2}(T+1) \rfloor / T)$
    and upper bound
Quantile$(\{\phi(f_1), \dots, \phi(f_T)\}; \lceil \left(1-\frac{\alpha}{2}\right)(T+1) \rceil / T)$
```

5.2 Bayesian Models with a Prior Sampler

In the previous section, we showed that one can get exact uncertainty sets for Bayesian models if an exact posteior sampler is available. However, for many Bayesian models, such as Bayesian neural networks (Bishop and Nasrabadi, 2006, chap 5.7) and latent Dirichlet allocation (Blei et al., 2003), it would be prohibitively expensive to sample from the exact posterior distribution. In such settings, practitioners often resort to approximating the posterior distribution, e.g., by using variational inference or Markov Chain Monte Carlo samplers (Bishop and Nasrabadi, 2006, ch. 10-11). However, when we only have access to samples from an approximate posterior distribution, it is not obvious how we can salvage our exact credible interval guarantee in Equation (10). In this section, we provide an algorithm that works without exact posterior samples, at the cost of providing weaker guarantees. Specifically, we guarantee validity with respect to the *prior* instead of the posterior. To do this, we recruit tools from conformal inference (See Section 2.)

Conformal inference is most often applied in frequentist settings, and allows one to construct prediction sets $C(x_i)$ for each new label y_i that enjoy finite-sample coverage guarantees. Specifically, the guarantee is that $\mathbb{P}\left(y \in C(\hat{f}(x))\right) \geq 1 - \alpha$, where x and y are random, and C is a random function of $\hat{f}(x)$ that also depends on held out calibration samples. Here, $\hat{f}$ is an arbitrary predictor for y that takes x as input. (Technically, there is an assumption that $\hat{f}$ treats the data symmetrically, but this is not important for our discussion here.) Conformal prediction requires T calibration samples for which both the prediction $\hat{f}(x)$ and the outcome y are observed.

We give a strategy for computing an uncertainty set $C_{\mathrm{conformal}}$ that is analogous to the conformal inference result, except that instead of giving an uncertainty set for a new label, we give an uncertainty set for the explanation of a model. The central challenge to applying conformal inference to our setting is obtaining our calibration samples; we usually do not know the true model explanation for any dataset. We get around this problem by sampling models i.i.d. from our prior distribution, $f_1, \ldots, f_t, \ldots, f_T \sim p(f)$. Recall that since our models are probabilistic, given an input x_i, we can sample a label $y_i^t \sim f_t(x_i)$ from the distribution predicted by the model. By pairing each original input x_i with the corresponding resampled label y_i^t, we have a dataset $D_t = \{(x_1, y_1^t), \ldots, (x_n, y_n^t)\}$ drawn from the model f_t. We can then train a model $\hat{f}_t = \mathcal{A}(D_t)$ on this new dataset. This gives us T examples where we can observe the ground truth explanation $\phi(f_t)$ and an estimated explanation $\phi(\hat{f}_t)$. We compare how close $\phi(f_t)$ and $\phi(\hat{f}_t)$ tend to be using a *nonconformity score*, such as the distance $\|\phi(f_t) - \phi(\hat{f}_t)\|$. These examples act as our calibration dataset in Algorithm 3.

Algorithm 3: CEI: Conformal Explanation Intervals

```
Input  : Model-fitting algorithm $\mathcal{A}$, dataset
         $D = (x_1, y_1), \ldots, (x_n, y_n)$
Input  : Nonconformity score $s : \Phi \times \Phi \to \mathbb{R}$
Train a model $\hat{f} = \mathcal{A}(D)$ using the dataset
Explain the trained model $\hat{\phi} = \phi(\hat{f})$
for $t = 1, \ldots, T$ do
    Sample a model $f_t \sim p(f)$
    Sample a dataset of labels $y_i^t \sim f_t(x_i)$
    Define the synthetic dataset
      $D_t = \{(x_1, y_i^t), \ldots, (x_n, y_n^t)\}$
    Train a model $\hat{f}_t = \mathcal{A}(D_t)$
    Explain the sampled model $\phi_t = \phi(f_t)$ and the
      trained model $\hat{\phi}_t = \phi(\hat{f}_t)$
    Compute the nonconformity score $s_t = s(\phi_t, \hat{\phi}_t)$
end
Set the threshold $\tau$ as the
  $\lceil (1-\alpha)(T+1) \rceil / T$-quantile of the set $\{s_1, \ldots, s_T\}$
Return: $C_{\mathrm{conformal}} = \{\varphi \in \Phi : s(\varphi, \phi(\hat{f})) \leq \tau\}$
```

The uncertainty set $C_{\mathrm{conformal}}$ has the following guarantee:

Proposition 2. *The confidence interval $C_{conformal}$ given by Algorithm 3 includes the model f with high probability over the prior distribution:*

$$\mathbb{P}_{f \sim p(f)}\left(\phi(f) \in C_{conformal}\right) \geq 1 - \alpha \tag{11}$$

Here, f is random due to the prior and $C_{\mathrm{conformal}}$ is random due to the data and the calibration samples. Note that computing $C_{\mathrm{conformal}}$ does not rely on us knowing the posterior distribution. We only need some algorithm $\mathcal{A}$, such as an empirical risk minimizer that, given a dataset, returns a model $\hat{f} \in \mathcal{F}$. However, this guarantee is weaker than the guarantee we got when we had access to the posterior distribution in Equation (10). Note that we are not conditioning on the data in Equation (11), and so $C_{\mathrm{conformal}}$ is not necessarily a credible interval under the posterior. Furthermore, by integrating over the dataset D, the (conditioned) guarantee in Equation (10) can be seen to imply the (unconditioned) guarantee in Equation (11). The condition in Proposition 2 is also satisfied by any credible interval C for the prior distribution. However, one should typically find that $C_{\mathrm{conformal}}$ can give much tighter uncertainty sets than this naive strategy since it is adaptive to the data. In this way, $C_{\mathrm{conformal}}$ can be viewed as an intermediate solution between a credible set for the prior and a credible set for the posterior; it is more adaptive to the data than the former but not as adaptive as the latter.

6 Experiments

We perform experiments on synthetic and real-world datasets. Synthetic datasets allow us to generate uncer-

	Coverage	Interval Width
Naive	0.5067 ± 0.002	0.2522 ± 0.001
Frequentist	1.0000 ± 0.000	3.9310 ± 0.003
Bayesian	0.9500 ± 0.001	0.9679 ± 0.001
Conformal	0.9600 ± 0.001	1.0493 ± 0.001

Table 2: A comparison of the coverage rate and interval width of the uncertainty sets from each method on synthetic data. Each of our proposed methods achieves the desired coverage of 0.95. The frequentist method is overly cautious due to its reliance on conservative learning theory results.

tainty sets for explanations where we know the true data generating distribution. This means that we can validate the coverage rates of our methods, which is difficult to do with real-world data where we do not know the data generating distribution. Experiments with real-world datasets give insight on how our methods scale to larger datasets and realistic distributions.

6.1 Experimental Setup

We perform four experiments under the Shapley value explainer. In a synthetic experiment, we first validate that the frequentist, Bayesian, and conformal methods all achieve the desired coverage rate and compare the size of the uncertainty sets each method gives. In the second synthetic experiment, we test the robustness of the Bayesian and conformal methods to violations of their assumptions by exploring settings where the prior distribution is misspecified. In the third experiment, we apply our methods to quantify the uncertainty of Deep SHAP explanations (Lundberg and Lee, 2017; Shrikumar et al., 2017) on the MNIST dataset. Lastly, we apply the conformal method to infer feature importance scores for a variety of real-world datasets.

Testing Coverage on Synthetic Data We consider a regression problem with three features. The data is generated according to the following distributions:

$$x_i \sim \mathcal{N}\left(\mu = \mathbf{0}, \Sigma = \begin{bmatrix} 1 & 0.99 & 0 \\ 0.99 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)$$

$$y_i = \theta^\top x + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1)$$

We fit linear models $f_\beta(x) = \mathcal{N}(\cdot\,; \mu = \beta^\top x, \sigma^2 = 1)$ that predict Gaussian distributions for y, where $\beta \in \mathbb{R}^3$. We independently sample $M = 100$ true models $\theta_1, \ldots, \theta_M \sim \mathcal{N}(\mathbf{0}, I_3)$, and for each true model θ_m, we sample a dataset D_θ consisting $n = 100$ examples. For the Bayesian and conformal methods, we the correct prior $\mathcal{N}(0, I_3)$ for θ. For the naive and conformal methods, we fit $T = 100$ linear models using Ridge regression. For the frequentist method, we use a standard uniform convergence result for linear regression. We use closed-form expression of Shapley value under the linear case. Refer to Appendix B for the detailed Rademacher complexity chosen and Shapley expression. For each method, we record the portion of the time that the uncertainty set includes the true explanation: coverage$(C) = \frac{1}{3M}\sum_{m=1}^{M}\sum_{i=1}^{n}\mathbb{1}\left\{\phi(\theta_m^{(i)}) \in C^{(i)}\right\}$. For each method, the targeted coverage rate is 0.95. We also report the average width of the uncertainty set for each method.

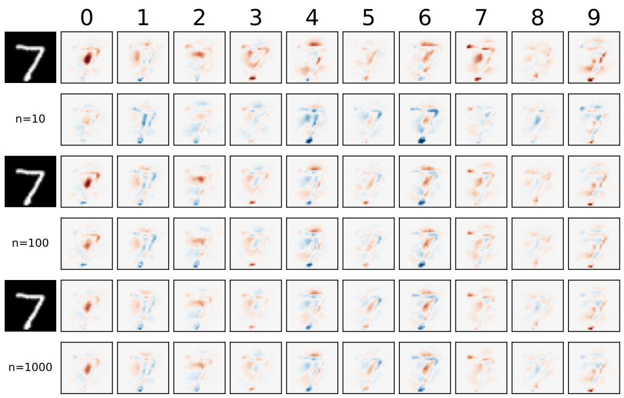

Figure 4: The range of plausible Deep SHAP feature attribution scores for MNIST digits. Here, the models were trained on 10, 100, and 1000 MNIST examples. All models explanations were computed with the same digit as input. Each column represents the impact of pixel values on the predicted probability assigned to that label (0-9). The top row next to each digit on the left is the upper bound for the feature attribution, and the bottom row is the lower bound. Red and blue represent positive and negative influence, respectively.

Testing Coverage under Model Misspecification We also explore how the Bayesian and conformal methods perform when the prior distribution is incorrect. We re-run the coverage experiment for the Bayesian and conformal methods, except with the true models sampled with a different mean $\mathcal{N}(\mathbf{1}, I_3)$ and with a different variance $\mathcal{N}\left(\mathbf{0}, \frac{1}{2}I_3\right)$. All other aspects of the experiment are unchanged.

MNIST Experiments We train models on the MNIST dataset, with the number of training examples varying between 10, 100, and 1000 examples. We train convolutional neural networks, so choosing a meaningful prior distribution for the model parameters is difficult. We explore using an isotropic Gaussian prior for all model parameters, and an empirical Bayes approach where we partially train an MNIST model on held-out data, then define a Gaussian prior centered at the weights of this trained model. We explain each model using the Deep SHAP explainer. We then generate conformal explanation intervals for each explanation. We investigate which pixels have high/low uncertainty, and how uncertainty varies as the number of training

		Coverage	Interval Width
Bayesian	well-specified	0.9500±0.001	0.9679±0.001
	wrong mean	0.7511±0.001	0.8620±0.001
	wrong var	0.9600±0.001	0.8720±0.001
Conformal	well-specified	0.9600±0.001	1.0493±0.001
	wrong mean	0.9867±0.001	1.0634±0.001
	wrong var	0.9800±0.001	1.0578±0.001

Table 3: A comparison of the Bayesian and conformal methods when their assumptions are not met, due to a misspecified prior distribution. The Bayesian method loses coverage when the mean is misspecified in this case. The conformal method becomes unnecessarily cautious, giving wider intervals than necessary.

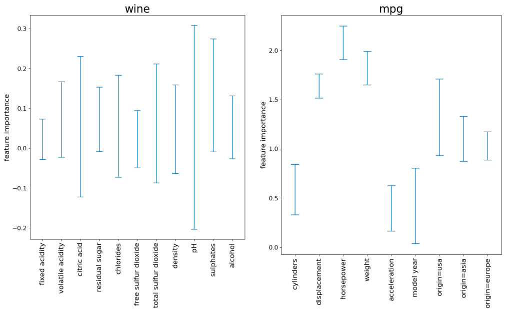

Figure 5: Feature importance scores (as measured by the mean Shapley value of each feature across the dataset) computed using the conformal method. For some datasets (e.g., MPG) there are significant differences between the importances assigned to different features. For other datasets (e.g., WINE), conclusions are more difficult to make. When features have overlapping feature importance uncertainty sets, it indicates that practitioners should be cautious when drawing conclusions.

examples increases.

Real-world Regression Experiments We consider eight tabular regression datasets: (all results except for WINE and MPG deferred to Appendix D). In each case, we train a neural network to predict a real-valued label. The model outputs a mean and variance for a Gaussian distribution, and is trained with the negative log-likelihood loss. The architecture has 2 hidden layers, each with 100 neurons, and uses ReLU activations. We compute uncertainty sets for the explanation of the true model using the conformal explanation intervals method. For the explanation, we use the average of the absolute value of the Shapley value of the feature across the dataset (a measure of feature importance). We set the prior distribution for each weight to be Gaussian with zero mean and variance as the reciprocal of the dimension of the layer. The prior for the biases are standard Gaussian distributions. We generate $T = 100$ calibration examples by sampling models from the prior.

6.2 Experimental Results

Testing Coverage on Synthetic Data We find that each of our proposed method achieves a coverage of at least 95% (the naive method has coverage close to 50%). See Table 2 for the complete results. The frequentist coverage interval tends to be overly conservative due to the worst-case perspective of uniform convergence, leading to a roughly 4x greater interval width when compared to the other methods, and 100% coverage in our experiments.

MNIST Experiments We find that the size of the dataset has an important impact on the degree of uncertainty in explanations. The uncertainty decreases rapidly as the dataset size increases (to a negligible value once there are 10,000 training examples). This could indicate that on MNIST, models tend to converge to similar optima (at least in terms of the explanations they admit) even with little data. This may also point to the importance of choosing reasonable priors when explaining neural networks, since our coverage guarantees are contingent on the chosen prior distribution. Additional results are included in Appendix D.

Real-world Regression Experiments We find that the strength of conclusions that can be drawn from the conformal method varies across datasets. For example, for the MPG dataset, the features `displacement`, `horsepower`, and `weight` have high importance with low uncertainty. However, in the PROTEIN dataset, it is difficult to make any meaningful conclusions about the relative importance of features, possibly due to the existence of competing models that use different features. Seven additional experimental results are included in Appendix D.

7 Discussion

We offer guidance to a practitioner deciding between the frequentist, Bayesian, and conformal approaches. If we do not have a prior for our model, the frequentist approach can give strong guarantees at the cost of large uncertainty sets. For Bayesian models, we recommend using the fully Bayesian approach when the posterior admits exact sampling, since this gives stronger guarantees. When an exact posterior sampler is not available, the conformal approach can recover a weaker guarantee that can still give tight uncertainty sets.

In this work we show how to give confidence sets for explanations of the data generating process. We caution, however, that explanations of the true data generating distribution do not in general have causal implications. Still, we hope that formalizing a connection between model explanations and the data generating distribution can help users understand which explanations are the result of spurious correlations, and which are meaningful.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Antorán, J., Bhatt, U., Adel, T., Weller, A., and Hernández-Lobato, J. M. (2020). Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*.

Awasthi, P., Frank, N., and Mohri, M. (2020). On the rademacher complexity of linear hypothesis sets. *arXiv preprint arXiv:2007.11045*.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020). An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. *Molecular Therapy-Nucleic Acids*, 22:362–372.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Black, E., Raghavan, M., and Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Carbonneau, R., Laframboise, K., and Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3):1140–1154.

Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. N. (2019). Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32.

Dong, J. and Rudin, C. (2019). Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*.

Fong, R. and Vedaldi, A. (2019). Explanations for attributing deep neural network predictions. In *Explainable ai: Interpreting, explaining and visualizing deep learning*, pages 149–167. Springer.

Frye, C., Rowat, C., and Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33.

Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Jabeur, S. B., Mefteh-Wali, S., and Viviani, J.-L. (2021). Forecasting gold price with the xgboost algorithm and shap interaction values. *Annals of Operations Research*, pages 1–21.

Janzing, D., Blöbaum, P., and Minorics, L. (2020a). Quantifying causal contribution via structure preserving interventions. *arXiv preprint arXiv:2007.00714*.

Janzing, D., Minorics, L., and Blöbaum, P. (2020b). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR.

Johnsen, P. V., Riemer-Sørensen, S., DeWan, A. T., Cahill, M. E., and Langaas, M. (2021). A new method for exploring gene–gene and gene–environment interactions in gwas with tree ensemble methods and shap values. *BMC bioinformatics*, 22(1):1–29.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Lakkaraju, H., Arsov, N., and Bastani, O. (2020). Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR.

Lee, E., Braines, D., Stiffler, M., Hudler, A., and Harborne, D. (2019). Developing the sensitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 349–356. SPIE.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Ley, D., Bhatt, U., and Weller, A. (2021). $\{\delta\}$-clue: Diverse sets of explanations for uncertainty estimates. *arXiv preprint arXiv:2104.06323*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Marx, C., Calmon, F., and Ustun, B. (2020). Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR.

Mokhtari, K. E., Higdon, B. P., and Başar, A. (2019). Interpreting financial time series with shap values. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pages 166–172.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Ning, Y., Ong, M. E. H., Chakraborty, B., Goldstein, B. A., Ting, D. S. W., Vaughan, R., and Liu, N. (2022). Shapley variable importance cloud for interpretable machine learning. *Patterns*, 3(4):100452.

Park, Y., Maddix, D., Aubet, F.-X., Kan, K., Gasthaus, J., and Wang, Y. (2022). Learning quantile functions without quantile crossing for distribution-free time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pages 8127–8150. PMLR.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Raccuglia, P., Elbert, K. C., Adler, P. D., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., and Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sagi, O. and Rokach, L. (2020). Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61:124–138.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shaikhina, T., Bhatt, U., Zhang, R., Georgatzis, K., Xiang, A., and Weller, A. (2021). Effects of uncertainty on the quality of feature importance explanations. In *AAAI Workshop on Explainable Agency in Artificial Intelligence*.

Shapley, L. (1953). A value for n-person games, contributions to the theory of games (kuhn, hw, tucker, aw, eds.), 307-317. *Annals of Mathematical Studies*, 28:275–293.

Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., and Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers & Operations Research*, 119:104926.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wen, X., Xie, Y., Wu, L., and Jiang, L. (2021). Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with lightgbm and shap. *Accident Analysis & Prevention*, 159:106261.

Zhou, X., Wen, H., Li, Z., Zhang, H., and Zhang, W. (2022). An interpretable model for the susceptibility of rainfall-induced shallow landslides based on shap and xgboost. *Geocarto International*, (just-accepted):1–27.

A Proofs

Lemma 1. *If uniform convergence holds, then with probability at least* $1-\alpha$*,*

$$\mathcal{L}_n(f^*) \leq \inf_{f \in \mathcal{F}} \mathcal{L}_n(f) + 2\epsilon_n.$$

Proof of Lemma 1. To see this, denote by the event $E = \{\sup_{f\in\mathcal{F}} |\mathcal{L}(f) - \mathcal{L}_n(f)| \leq \epsilon\}$, which occurs with probability at least $1-\alpha$. If the event E occurs, then for all models $f \in \mathcal{F}$ we have

$$\begin{aligned}
\mathcal{L}_n(f^*) &\leq \mathcal{L}(f^*) + \epsilon_n && (E \text{ occurred}) \\
&\leq \mathcal{L}(f) + \epsilon_n && (\text{Optimality of } f^*) \\
&\leq \mathcal{L}_n(f) + 2\epsilon_n. && (E \text{ occurred})
\end{aligned}$$

Since this inequality holds for all $f \in \mathcal{F}$, it also holds for the infimum over $\mathcal{F}$. This gives the result in Equation (4). □

B Example: Linear Regression

We will walk through an example in which we infer the Shapley values of an unknown linear function.

B.1 Shapley Value Derivation for Linear Models

Consider a linear model given by $f_\theta(x) = \theta^\top x$, where $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$. We use the capital X to represent the random variable for the feature vector, and the lower case x to represent a fixed value for this random variable. Given a coalition of features $S \subseteq \{1, \ldots, d\}$, we can write the feature vector x as $x = [x_S, x_{\overline{S}}]$ where $x_S = \{x_i : i \in S\}$ and $x_{\overline{S}} = \{x_i : i \notin S\}$. The prediction associated with the coalition x_S can be defined as

$$\begin{aligned}
f_\theta(x_S) &:= \mathbb{E}\left[f_\theta([x_S, X_{\overline{S}}])\right] && (12) \\
&= \mathbb{E}\left[\theta^\top [x_S, X_{\overline{S}}]\right] && (13) \\
&= \theta_S^\top x_S + \theta_{\overline{S}}^\top \mathbb{E}\left[X_{\overline{S}}\right] && (14) \\
&= \theta_S^\top x_S + \theta_{\overline{S}}^\top x_{\overline{S}} + \theta_{\overline{S}}^\top \mathbb{E}\left[X_{\overline{S}} - x_{\overline{S}}\right] && (15) \\
&= f_\theta(x) + \sum_{i \notin S} \theta_i \mathbb{E}\left[X_i - x_i\right] && (16)
\end{aligned}$$

The Shapley value for a linear model on a particular instance x can then be written as:

$$\begin{aligned}
\phi_i^x(f_\theta) &= \sum_{S \subseteq [d]\backslash i} \frac{|S|!(d-|S|-1)!}{d!} \left(f_\theta(x_{S\cup\{i\}}) - f_\theta(x_S)\right) && (17) \\
&= \sum_{S \subseteq [d]\backslash i} \frac{|S|!(d-|S|-1)!}{d!} \left(\left(f_\theta(x) + \sum_{j \notin S\cup\{i\}} \theta_j \mathbb{E}\left[X_j - x_j\right] \right) - \left(f_\theta(x) + \sum_{j \notin S} \theta_j \mathbb{E}\left[X_j - x_j\right] \right) \right) && (18) \\
&= \sum_{S \subseteq [d]\backslash i} \frac{|S|!(d-|S|-1)!}{d!} \theta_i \mathbb{E}\left[X_i - x_i\right] && (19) \\
&= \frac{1}{Z} \theta_i \mathbb{E}\left[X_i - x_i\right] && (20)
\end{aligned}$$

where $Z = \left(\sum_{S\subseteq[d]\backslash i} \frac{|S|!(d-|S|-1)!}{d!}\right)^{-1}$ is the normalizing constant. Using the efficiency property of the Shapley value which states that $\sum_{i\in[d]} \phi_i^x(f_\theta) = f_\theta(x)$, we can compute the normalizing constant Z by noting that:

$$f_\theta(x) = \sum_{i\in[d]} \phi_i^x(f_\theta) \tag{21}$$

$$= \frac{1}{Z}\sum_{i\in[d]} \theta_i \mathbb{E}\left[X_i - x_i\right] \tag{22}$$

giving us that

$$Z = \frac{1}{f_\theta(x)}\sum_{i\in[d]} \theta_i \mathbb{E}\left[X_i - x_i\right] \tag{23}$$

We can then write the Shapley value as

$$\phi_i^x(f_\theta) = \frac{f_\theta(x)}{\theta^\top(\mathbb{E}\left[X\right] - x)}\theta_i \mathbb{E}\left[X_i - x_i\right] \tag{24}$$

when the denominator is nonzero, and $\phi_i^x(f_\theta) = 0$ when the denominator is equal to zero.

B.2 Uniform Convergence for Squared Loss of Linear Models

Uniform convergence results bound (with high probability) the disagreement $\sup_{f\in\mathcal{F}} |\mathcal{L}(f) - \mathcal{L}_n(f)|$ between the sample loss and population loss over a model class. Uniform convergence results are often stated in terms of the Vapnik–Chervonenkis dimension, Rademacher complexity, Gaussian complexity, covering number, and other notions of complexity of the model class. Below, we display a few standard results that, together, give a uniform convergence result for linear models with squared error loss.

Theorem 1 (Awasthi et al. (2020))**.** *Let $\mathcal{F} = \left\{x \mapsto \theta^\top x : \|\theta\|_p \leq w\right\}$ be a family of linear functions defined over $\mathbb{R}^d$ with bounded weight in ℓ_p-norm. Then, the empirical Rademacher complexity of $\mathcal{F}$ for a sample $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ admits the following upper bounds:*

$$\hat{\mathcal{R}}_S\left(\mathcal{F}\right) \leq \frac{w}{n}\|\mathbf{X}\|_{\mathrm{Fr}}$$

where $\mathbf{X}$ is the $d \times n$-matrix with $\mathbf{x}_i$s as columns: $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_n]$.

Theorem 2. *For the squared error loss $\ell(y, \hat{y}) = (y - \hat{y})^2$, let $\ell \circ \mathcal{F} := \left\{x \mapsto (f_\theta(x) - f^*(x))^2 : f_\theta \in \mathcal{F}\right\}$. Assume that $\sup_{x\in\mathcal{X}, f_\theta\in\mathcal{F}} \left(f_\theta(x) - f^*(x)\right)^2 \leq M^2$. Then for any sample $S = \{x_1, \ldots, x_n\}$,*

$$\hat{\mathcal{R}}_S(\ell \circ \mathcal{F}) \leq 2M\hat{\mathcal{R}}_S(\mathcal{F}) \tag{25}$$

Theorem 3. *With probability at least $1 - \delta$, for all $f \in \mathcal{F}$ and distributions over $\mathcal{X} \times \mathcal{Y}$ the following holds:*

$$\mathcal{L}(f) - \mathcal{L}_n(f) \leq 2\hat{\mathcal{R}}_S(\ell \circ \mathcal{F}) + 3\sqrt{\frac{\ln 1/\delta}{2n}} \tag{26}$$

Together, Theorems 1-3 give us the following result for linear models with the squared error loss. For all $f \in \mathcal{F}$ and any distribution over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$,

$$\mathcal{L}(f) - \mathcal{L}_n(f) \leq \frac{4Mw}{n}\|\mathbf{X}\|_{\mathrm{Fr}} + 3\sqrt{\frac{\ln 1/\delta}{2n}} \tag{27}$$

where $\mathbf{X}, M, w, n, \delta$ are defined as in Theorems 1-3.

C Conformal Prediction

Our Algorithm 3 is heavily inspired by conformal prediction, a simple and effective method for constructing statistically rigorous uncertainty intervals (Angelopoulos and Bates, 2021; Lei and Wasserman, 2014; Vovk et al., 2005). In the standard conformal prediction setup, we have an i.i.d. calibration dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ and some new input X_{test} for which we want to predict the label Y_{test}. Given a black-box model $\hat{f}$ trained (on separate data) to predict the label, we want to construct uncertainty estimates for the predictions made by the black-box model. Conformal prediction tells us how to construct an uncertainty interval C_{test} such that the ground truth value Y_{test} is included in the uncertainty interval with some chosen probability $1 - \alpha$, such at 95%:

$$P(Y_{test} \in C_{\text{test}}) \geq 1 - \alpha \tag{28}$$

Perhaps the biggest advantage of conformal prediction is that it applies under extremely weak assumptions. As long as the model-fitting algorithm $\mathcal{A}$ treats the data symmetrically (e.g., a time series forecast that weighs recent data more heavily does not treat data symmetrically) and the data is i.i.d. (in fact, the weaker condition of exchangeability is sufficient), the uncertainty interval will be valid. No additional distributional assumptions on the data generating process are needed.

The central component of a conformal prediction algorithm is the nonconformity score. The nonconformity score $s(Y, \hat{f}(X))$ evaluates the disagreement between the observed outcome and the model's prediction. For a binary classifier that outputs a probability $\hat{p}(X_{\text{test}})$ that $Y_{\text{test}} = 1$, a reasonable nonconformity score would be $s(Y_{\text{test}}, \hat{p}(X_{\text{test}})) = |Y_{\text{test}} - \hat{p}(X_{\text{test}})|$. For a regression model, a reasonable nonconformity score would be $s(Y_{\text{test}}, \hat{\mu}(X_{\text{test}})) = |Y_{\text{test}} - \hat{\mu}(X_{\text{test}})|$ where $\hat{\mu}(X_{\text{test}})$ is the predicted mean.

Importantly for our purposes, conformal prediction requires us to have i.i.d. examples to calibrate our uncertainties. When we want an uncertainty interval for an outcome Y_{test}, this is not a problem since we often have access to pairs of true outcomes Y_i and predicted outcomes $\hat{f}(X_i)$. However, in our case we want an uncertainty interval for the explanation $\phi(f)$. Problematically, we never observe a true explanation because we never observe i.i.d. examples of a "true explanation" $\phi(f)$ and an estimated explanation $\phi(\hat{f})$. Usually, we only observe a single dataset generated from a single model f. However, if we have a prior distribution for the model, then we can simulate i.i.d. examples of true explanations $\phi(f)$ and estimated explanations $\phi(\hat{f})$ then run conformal prediction.

D Additional Experimental Results

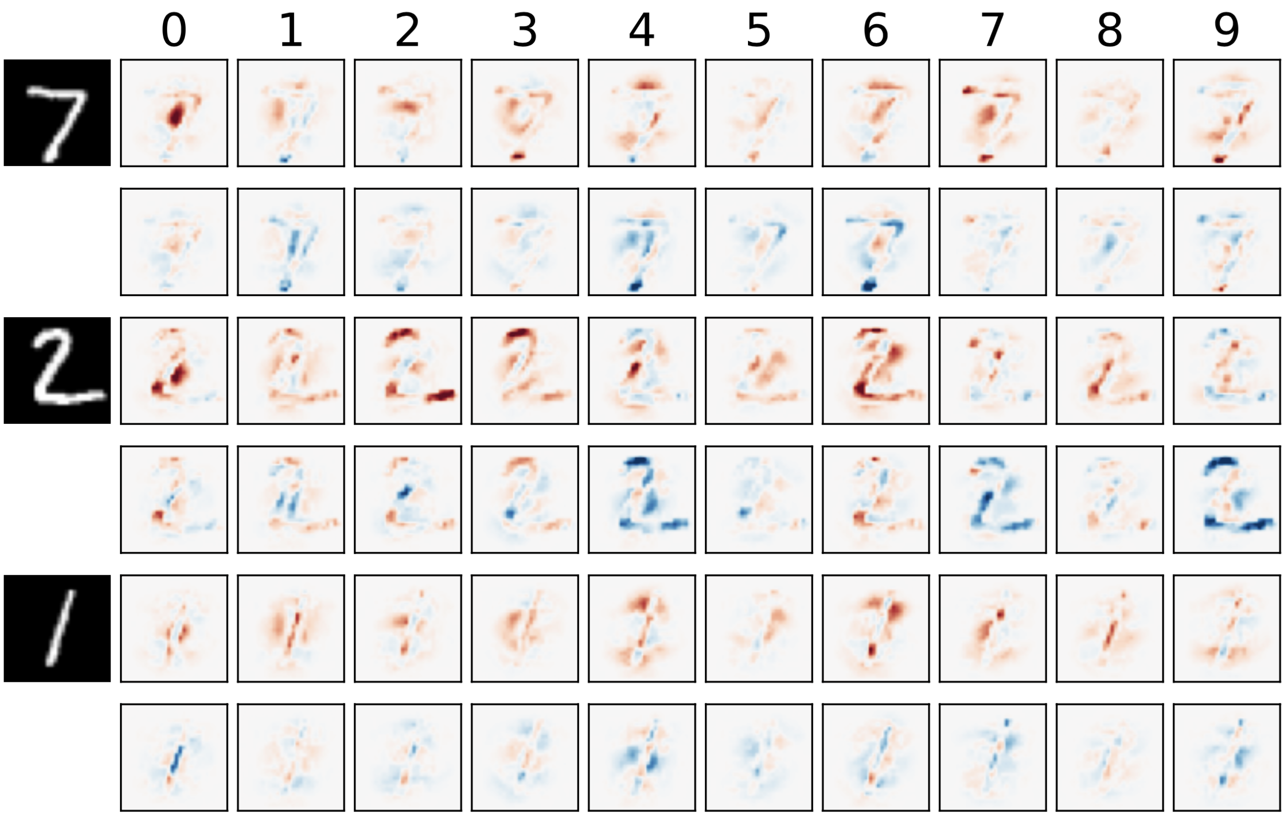

Figure 6: Additional results for Deep SHAP explanations. Here, all models were trained on 100 MNIST examples. Each column represents the impact of pixel values on the predicted probability assigned to that label (0-9). The top row next to each digit on the left is the upper bound for the feature attribution, and the bottom row is the lower bound. Red and blue represent positive and negative influence, respectively.

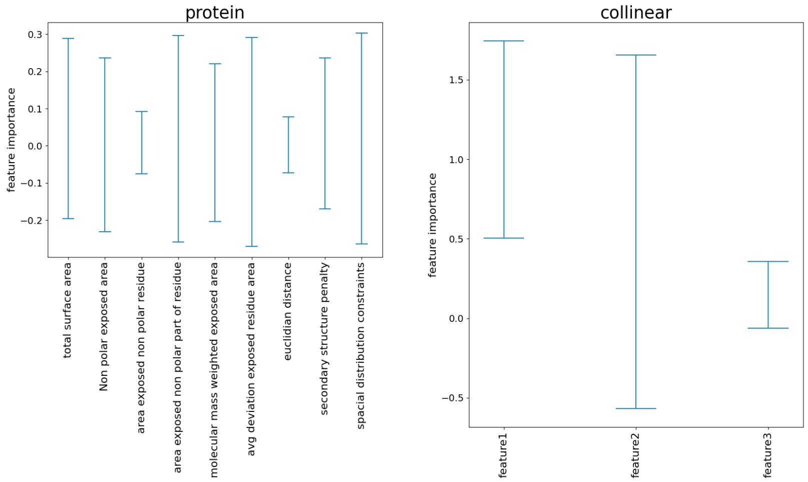

Figure 7: Feature importance scores (as measured by the mean Shapley value of each feature across the dataset). Confidence intervals are computed using the conformal explanation intervals method.

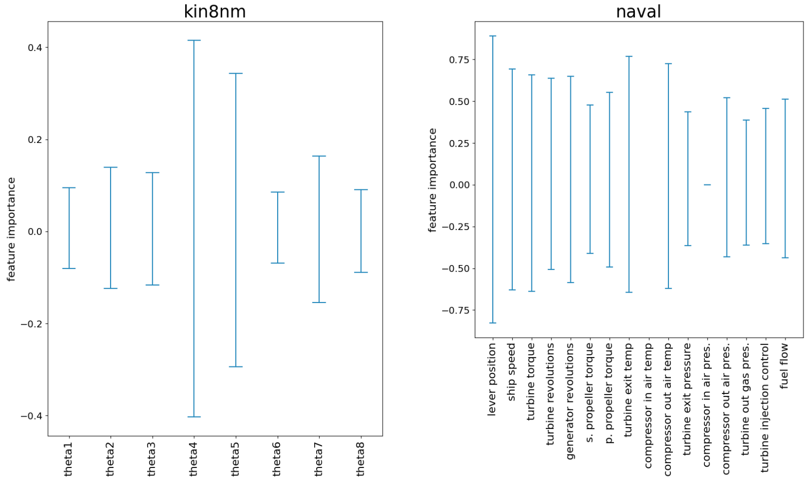

Figure 8: Feature importance scores (as measured by the mean Shapley value of each feature across the dataset). Confidence intervals are computed using the conformal explanation intervals method.

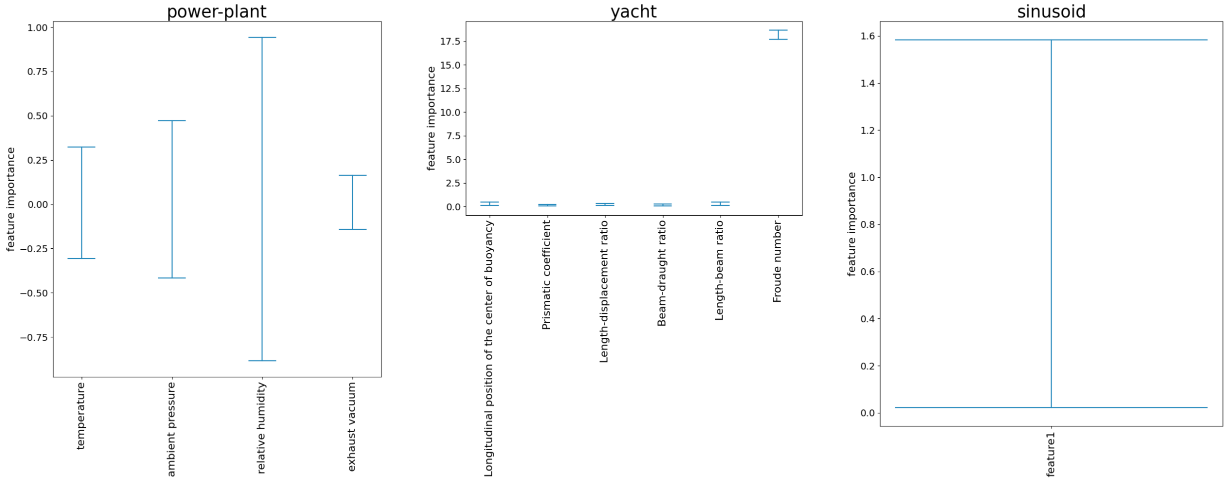

Figure 9: Feature importance scores (as measured by the mean Shapley value of each feature across the dataset). Confidence intervals are computed using the conformal explanation intervals method.