Identifying Drivers of Predictive Aleatoric Uncertainty

Pascal Iversen[1,2], **Simon Witzke**[1], **Katharina Baum**[1,2,3] and **Bernhard Y. Renard**[1,2,3]
[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Germany
[2]Freie Universität Berlin, Department of Mathematics and Computer Science, Berlin, Germany
[3]Windreich Department of Artificial Intelligence and Human Health & Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA
katharina.baum@fu-berlin.de, bernhard.renard@hpi.de

Abstract

Explainability and uncertainty quantification are key to trustable artificial intelligence. However, the reasoning behind uncertainty estimates is generally left unexplained. Identifying the drivers of uncertainty complements explanations of point predictions in recognizing model limitations and enhancing transparent decision-making. So far, explanations of uncertainties have been rarely studied. The few exceptions rely on Bayesian neural networks or technically intricate approaches, such as auxiliary generative models, thereby hindering their broad adoption. We propose a straightforward approach to explain predictive aleatoric uncertainties. We estimate uncertainty in regression as predictive variance by adapting a neural network with a Gaussian output distribution. Subsequently, we apply out-of-the-box explainers to the model's variance output. This approach can explain uncertainty influences more reliably than complex published approaches, which we demonstrate in a synthetic setting with a known data-generating process. We substantiate our findings with a nuanced, quantitative benchmark including synthetic and real, tabular and image datasets. For this, we adapt metrics from conventional XAI research to uncertainty explanations. Overall, the proposed method explains uncertainty estimates with little modifications to the model architecture and outperforms more intricate methods in most settings.

1 Introduction

Uncertainty quantification and explainability are crucial for adopting machine learning (ML) systems in safety-critical applications, ensuring trust, reliability, and fairness [Abdar *et al.*, 2021; Vilone and Longo, 2020; Lötsch *et al.*, 2022]. Predictive uncertainty in ML refers to the degree of confidence associated with a model's predictions [Chua *et al.*, 2023]. It can be decomposed into an epistemic and aleatoric component [Kendall and Gal, 2017]. Epistemic uncertainty stems from data scarcity, such as underrepresented conditions, covariate shift, and model misspecification and can generally be reduced with more data. Aleatoric uncertainty, arising from the random error of the true relationship between inputs, and targets. It reflects irreducible variability in the data. Uncertainty estimation is critical in risk management. It allows taking conservative action, relying on the model only when it exhibits high confidence in its predictions [Kompa *et al.*, 2021].

Explainability encompasses methods that enhance the transparency of ML models by highlighting how features influence model output or by rendering the internal computations of black-box models more interpretable. Explainability methods enable understanding whether a model has learned relevant patterns from the input data and can reveal interesting, previously unknown associations [Samek *et al.*, 2021; Schwalbe and Finzel, 2023]. Uncertainty quantification and explainability ensure accountable, informed, and responsible decision-making and help mitigate biases and risks [Bhatt *et al.*, 2021; McGrath *et al.*, 2023].

In most applications, explainability focuses on interpreting point predictions [Vilone and Longo, 2020]. There is a significant gap in understanding and explaining the drivers of uncertainty estimates. When an ML algorithm is deployed and yields a substantial uncertainty estimate for a specific instance, the possible courses of action involve abstaining from employing the model if alternatives are available or accepting the increased risk. With explainable uncertainties, users gain the capability to identify the factors contributing to elevated uncertainty levels. This understanding allows domain experts to judge their relevance in a given scenario. Additionally, it provides valuable insights into modifications required to augment the model's predictive certainty and performance. In cases where abstaining from model usage is still necessary, factors influencing the decision can be understood and communicated. For example, if such an uncertainty factor is a feature indicating a person's age, it could point to an issue where the model's predictions are more uncertain for specific age groups, even if the age distribution is balanced in the training data. This effect would be undetectable by naive explanations. While detecting and explaining distribution shifts and epistemic uncertainty is an equally interesting problem [Brown and Talbert, 2022], we focus our work on aleatoric uncertainty. Aleatoric uncertainty estimates and their explanations are relevant for domains where the noise of the outcome of interest is not constant across independent variables, i.e., heteroscedastic settings. In these cases, aleatoric uncer-

tainty explanations offer complementary information to explanations of point predictions, as the relevant variables influencing mean and variance might differ significantly. Heteroscedastic settings emerge, for example, in the estimation of biophysical variables [Lázaro-Gredilla *et al.*, 2014], the estimation of cosmological redshifts [Almosallam *et al.*, 2016], and robotics and vehicle control [Bauza and Rodriguez, 2017; Smith *et al.*, 2018; Liu *et al.*, 2021].

Explanations can be categorized as either local (instance-based) or global (across the whole input space) [Schwalbe and Finzel, 2023; Adadi and Berrada, 2018]. A local explanation of the model's uncertainty could foster more transparent discussions about ML-assisted decisions and risks, increasing trust. Global explanations serve to detect general drivers of uncertainty and certainty. These can then be leveraged to formulate hypotheses to improve the model or to detect unintended shortcuts in the uncertainty estimation process, such as spurious correlations or biases.

There is little prior work on explaining uncertainties, and existing literature mainly focuses on classification tasks and generally relies on Bayesian neural networks (BNNs) or technical intricacies such as auxiliary generative models [Antoran *et al.*, 2021; Perez *et al.*, 2022; Ley *et al.*, 2022; Wang *et al.*, 2023]. BNNs assign probability distributions to network weights to capture uncertainty [MacKay, 1992]. However, due to their computational complexity and involved training process, BNNs have not been as widely adopted as classical neural networks [Lakshminarayanan *et al.*, 2017].

We propose a straightforward and scalable approach for explaining uncertainties in a heteroscedastic regression setting that can be readily integrated into ML pipelines (see Figure 1). We extend point prediction models to additionally estimate parameters of the spread of a given probability distribution. Specifically, we predict parameters of a Gaussian distribution as in a heteroscedastic regression model [Bishop, 1994]. The variance parameter of the Gaussian can be interpreted as a measure of the aleatoric uncertainty of the model. We can then use any explainability method to explain the variance estimate provided by this distributional model. By highlighting input features contributing to the variance output, we identify the inputs contributing to model uncertainty.

Currently, there is a gap in the comparative evaluation of uncertainty explainers in the literature. Therefore, we introduce a benchmark with synthetic data with a known data-generating process to analyze a method's ability to detect uncertainty drivers. In addition, we introduce MNIST+U, an image dataset including known uncertainty drivers based on MNIST [Deng, 2012]. We compare our approach to Counterfactual Latent Uncertainty Explanations (CLUE) [Antoran *et al.*, 2021] and InfoSHAP [Watson *et al.*, 2023]. For this purpose, we adapt unsupervised XAI metrics to evaluate the uncertainty explainers.

In summary, our contribution is as follows: We propose a straightforward explanation method for uncertainty and evaluate it against existing approaches. Further, we devise tabular and image benchmarks, including established metrics from the XAI field. Thereby, we provide a resource for informed usage of uncertainty explanation methods.

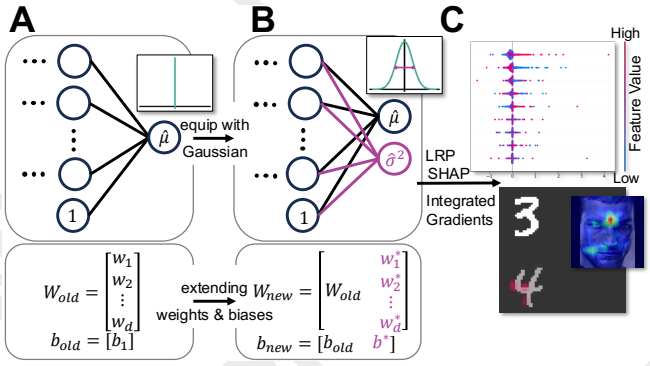

Figure 1: Overview of the variance feature attribution pipeline. (A) A point prediction model with an output layer with weight matrix $\boldsymbol{W}_{old} \in \mathbb{R}^{d\times 1}$ and a scalar bias. We equip this model with a Gaussian distribution resulting in (B), a model with output weight matrix $\boldsymbol{W}_{new} \in \mathbb{R}^{d\times 2}$ and bias $\boldsymbol{b}_{new} \in \mathbb{R}^2$. The two outputs are the mean $\hat{\mu}$ and the variance $\hat{\sigma}^2$ of the predictive distribution. (C) From there, we can explain the variance using any suitable explainability method, resulting in attributions to the input features that can be used to understand the drivers of the model's aleatoric uncertainty.

1.1 Related Work

In some research communities, such as causal inference, graphical models, and Gaussian processes, explicitly modeling uncertainty is a prominent area of interest. Furthermore, uncertainty quantification and explainability are rich areas of research within the deep learning field [Abdar *et al.*, 2021; Vilone and Longo, 2020]. Yet, few researchers have recognized the importance of explaining the sources of uncertainty in deep learning predictions. Yang and Li [2023] have developed an explainable uncertainty quantification approach for predicting molecular properties. They employ message-passing neural networks and generate unique uncertainty distributions for each atom of a molecule. This approach is inherently specialized for graph-based representations of molecules. CLUE [Antoran *et al.*, 2021] and related approaches [Perez *et al.*, 2022; Ley *et al.*, 2022] derive counterfactual explanations by optimizing for an adversarial input that is close to the original input but minimizes uncertainty. The adversarial input is constrained to the data manifold with a deep generative model of the input data to prevent out-of-distribution explanations. This requires an optimization process for each instance's explanation and the training of an auxiliary generative model, rendering CLUE and its extensions computationally demanding and difficult to implement. Additionally, Antoran *et al.* [2021] developed an evaluation method for contrastive explanations of uncertainty. Wang *et al.* [2023] have developed a gradient-based uncertainty attribution method for image classification with BNNs. They modify the backpropagation to attain complete, non-negative pixel attribution. To detect and explain model deterioration, Mougan and Nielsen [2023] use classical ML methods and bootstrapping. They train a model and obtain uncertainty estimates on a test set transformed with an artificial distribution shift. In a second step, they train another model to predict the uncertainty estimates from the first step. Subsequently, Shapley values are estimated for the second model to explain the

uncertainty. Mehdiyev *et al.* [2023] employ quantile regression forests to obtain prediction intervals that quantify uncertainty. They extract feature attributions for the uncertainty by estimating Shapley values directly for these prediction intervals as output. Watson *et al.* [2023] introduce variants of the Shapley value algorithm to explain higher moments of the predictive distribution by quantifying feature contributions to conditional entropy. They use a split conformal inference strategy. They first train a base model to predict conditional probabilities. Subsequently, they fit an auxiliary model to the base model's log square residuals. They interpret the estimated Shapley values of this residual model as uncertainty explanations. Bley *et al.* [2025] propose a second-order uncertainty attribution method that explains predictive uncertainty by computing the covariance of first-order feature attributions across model ensembles. While this approach offers insights into individual and joint feature contributions to uncertainty, it is limited to ensembles and cannot explain aleatoric uncertainty.

2 Methods

2.1 Deep Heteroscedastic Regression and Extension of Pre-trained Models

We explain uncertainties in neural network regressors using deep distributional networks. Specifically, we employ heteroscedastic regression with a Gaussian output to model variance in addition to the mean and, therefore, capture input dependence of the output noise. Here, we consider a regression setting with n independent training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with input feature vector $\boldsymbol{x}_i \in \mathbb{R}^k$ and target $y_i \in \mathbb{R}$, $i = 1, \ldots, n$. Instead of providing a complete picture of the conditional distribution of the target, deep regression models usually only estimate its conditional mean by optimizing the mean squared error (MSE) or comparable loss functions. In contrast, we assume a heteroscedastic Gaussian as the conditional distribution $y \mid \boldsymbol{x} \sim \mathcal{N}\left(\mu_{\boldsymbol{x}}, \sigma^2_{\boldsymbol{x}}\right)$ and represent its mean $\mu_{\boldsymbol{x}}$ and variance $\sigma^2_{\boldsymbol{x}}$ using a neural network $f_\theta : \mathbb{R}^k \to \mathbb{R} \times \mathbb{R}^+$ with weights $\boldsymbol{\theta}$ and two output neurons producing the mean and variance estimates $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = (\hat{\mu}_{\boldsymbol{x}}, \hat{\sigma}^2_{\boldsymbol{x}})$, respectively. As first described by Bishop [1994], we can then optimize the Gaussian negative log-likelihood (GNLL): $\mathcal{L} \propto \sum_{i=1}^n \left(\log(\hat{\sigma}^2_{\boldsymbol{x}_i}) + \frac{(y_i - \hat{\mu}_{\boldsymbol{x}_i})^2}{\hat{\sigma}^2_{\boldsymbol{x}_i}}\right)$ and interpret the predicted variance as a measure of the aleatoric uncertainty of the model. However, naively optimizing this criterion with overparametrized models such as deep neural networks can be unstable [Wong-Toi *et al.*, 2023; Nix and Weigend, 1994; Seitzer *et al.*, 2022]. In practice, these convergence difficulties can be mitigated by initially training the model using solely the MSE $\sum_{i=1}^n \left(y_i - \hat{\mu}_{\boldsymbol{x}_i}\right)^2$ and subsequently switching to the GNLL [Sluijterman *et al.*, 2023].

The two-stage training process aligns with transfer learning: MSE-based initial training serves as pre-training, followed by fine-tuning with the GNLL to capture predictive uncertainty. Extending existing pre-trained models to capture uncertainty is relevant when the model size and associated training costs make full re-training unfeasible. Pre-trained regression models can be extended by concatenating a column of randomly initialized weights to the weight matrix of the output layer to attain a variance estimate (see Figure 1).

2.2 Post-hoc Explanation of Predictive Variance

Classic explainability methods explain the predicted class or point prediction. In contrast, we want to explain the variance output in a heteroscedastic regression model. In these models, variance is an additional output to which we can apply any existing, appropriate explainability method. In principle, an uncertainty explanation can be achieved for any parametrized output distribution for which an explicit formulation of the uncertainty is available. In the case of a Gaussian output distribution, the application is most intuitive since its variance parameter is a direct output of the neural network. Furthermore, unlike distributions such as the Poisson or exponential distributions, the variance is uncoupled from the mean output. For binary classification, where the model outputs the parameter of a Bernoulli distribution, an entropy formulation of uncertainty can be utilized. Alternatively, explaining aleatoric uncertainty for classification can be approached by operating in the logit space.

We employ model-agnostic and model-specific post-hoc explainability methods to explain uncertainty. Model-specific methods are limited in the type of models that they can explain but may offer advantages such as lower computational complexity. In contrast, model-agnostic methods can be applied to any model [Adadi and Berrada, 2018]. For our experiments, we combine the approach described in Section 2.1 with multiple explainability methods and refer to this conjunction as Variance Feature Attribution (VFA) flavors. As the first explainability method, we use KernelSHAP [Lundberg and Lee, 2017], a model-agnostic, local explainability method. KernelSHAP approximates Shapley values using a weighted linear surrogate model with an appropriate weighting kernel (VFA-SHAP). Additionally, for image tasks, we leverage DeepSHAP [Lundberg and Lee, 2017], a variant of SHAP tailored for deep learning models, which combines SHAP values with the DeepLIFT algorithm [Shrikumar *et al.*, 2017] to efficiently attribute model predictions to input features. We also employ Integrated Gradients (IG) [Sundararajan *et al.*, 2017], which is a local, model-specific method and assigns feature importance by integrating predictions over a straight path from a baseline to the input (VFA-IG). Further, we use Layer-Wise Relevance Propagation (LRP) [Bach *et al.*, 2015], a local, model-specific explainability method developed for neural networks where the importance is distributed backward to the input layer by layer weighted by a neuron's contribution (VFA-LRP). We compare the VFA flavors to CLUE, for which we have to train a variational autoencoder on the train data and apply the optimization as detailed by Antoran *et al.* [2021]. CLUE attributions are the absolute differences between counterfactual and input feature vectors. CLUE is local and model-specific. Further, we reimplement InfoSHAP for regression, which estimates the uncertainty using an auxiliary model trained on the log-square residuals of a base model. The uncertainty attribution is attained by estimating the Shapley values of the auxiliary model [Watson *et al.*, 2023]. As InfoSHAP builds on SHAP, it is a model-agnostic, local explainability method. Global explanations

are obtained by averaging local method results over a dataset.

2.3 Uncertainty Explanation Evaluation Metrics

There is little prior work on evaluating the quality and properties of uncertainty explanations. Generally, high-quality explanations have to be robust, faithful, and highlight relevant input features. We extend established metrics for general XAI to the explanation of model uncertainty.

In a situation where ground truth noise drivers are known, we can examine if explanation methods correctly rediscover them. Arras *et al.* [2022] introduce metrics for this setting for classical XAI: *Relevance Rank Accuracy* (RRA) describes the proportion of known relevant features that are rediscovered by the explanation method for a given sample $\boldsymbol{x}_i$.

Relevance Mass Accuracy (RMA) describes the amount of relevance that is assigned to the ground truth features, normalized by the total amount of relevance. For uncertainty explanations, we judge if a method discovers features that correlate with the standard deviation of the target's heteroscedastic noise. To scrutinize global explanations, we apply these accuracy metrics to global feature attributions, giving rise to global relevance rank accuracy (GRA) and global relevance mass accuracy (GMA). Global accuracy measures how effectively a model detects general drivers of uncertainty across the entire dataset. In contrast, local accuracy indicates the model's ability to identify uncertainty sources for individual instances and is, therefore, a stricter criterion.

Alvarez-Melis and Jaakkola [2018] argue that *Robustness* is a key property of explanations, demanding that proximal inputs lead to similar explanations. They propose to evaluate robustness with local Lipschitz continuity: $\hat{L}(\boldsymbol{x}_i) = \max_{\boldsymbol{x}_j \in \mathcal{N}_\epsilon(\boldsymbol{x}_i)} \frac{\|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|_2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}$, where f is the explanation method. For a dataset with only continuous features, the perturbation space $\mathcal{N}_\epsilon(\boldsymbol{x}_i)$ is a ball with radius ϵ around sample $\boldsymbol{x}_i$. However, continuous perturbations lack meaning for categorical features. Instead, the perturbation space is defined as the set of data points close to $\boldsymbol{x}_i$: $\mathcal{N}_\epsilon(\boldsymbol{x}_i) = \{\boldsymbol{x}_j \in \mathcal{X} \mid \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \epsilon, \boldsymbol{x}_i \neq \boldsymbol{x}_j\}$, where $\mathcal{X}$ is the set of test inputs. Low Lipschitz estimates indicate small changes in the explanation upon perturbation and, therefore, high robustness. This notion of robustness can be extended to uncertainty explanations by applying it to the variance head predictions or an auxiliary uncertainty model.

Further, we analyze *Faithfulness* of the explanations. If an explanation is faithful, changing input features that are considered relevant should lead to a significant reduction in prediction performance. Commonly, this is measured as the increase of the loss upon perturbation of relevant features [Arras *et al.*, 2022]. However, the GNLL loss we use during training is a function of the mean and variance, and its magnitude is not interpretable. We aim to evaluate the perturbation's impact on the quality of the uncertainty estimate. Naturally, we demand that a higher uncertainty estimate should relate to a higher expected squared error of the mean prediction. Therefore, we measure the correlation between the squared residuals and the uncertainty estimates. Let $\mathbf{y} \in \mathbb{R}^n$ be the vector of ground truth target values from the test set, and let $\hat{\boldsymbol{\mu}}(\mathbf{X}) \in \mathbb{R}^n$ and $\hat{\boldsymbol{\sigma}}^2(\mathbf{X}) \in \mathbb{R}^n$ denote the predicted means and variances, respectively, for test inputs $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the number of test samples and d is the number of features. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be the feature attribution matrix, where $\mathbf{A}_{ij}$ is attribution of feature j for sample i, as produced by an uncertainty explainer applied to $\hat{\boldsymbol{\sigma}}^2(\mathbf{X})$. We first calculate the Spearman correlation $\rho_s = \text{corr}_s\left((\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}))^2, \hat{\boldsymbol{\sigma}}^2(\boldsymbol{X})\right)$. We then determine the k globally most important uncertainty features. Precisely, we compute the index set of the most important features $\mathcal{I}_k = \text{top-}k_j\left(\frac{1}{n}\sum_{i=1}^{n} |\mathbf{A}_{ij}|\right)$, i.e., the indices of the k features with the highest average absolute attribution values. Subsequently, we define a perturbed input matrix $\mathbf{X}' \in \mathbb{R}^{n \times d}$ by adding Gaussian noise to the $k = 3$ most important features for uncertainty prediction

$$\mathbf{X'}_{ij} = \begin{cases} \mathbf{X}_{ij} + \delta_{ij}, & \text{if } j \in \mathcal{I}_k \\ \mathbf{X}_{ij}, & \text{otherwise} \end{cases}, \quad \delta_{ij} \sim \mathcal{N}(0, 1).$$

Based on this, we calculate the correlation of the original residuals with the perturbed uncertainties $\rho'_s = \text{corr}_s\left((\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{X}))^2, \hat{\boldsymbol{\sigma}}^2(\boldsymbol{X}')\right)$ and expect the variance to be less expressive after the perturbation, i.e., the change $\rho'_s - \rho_s$ is negative, if the uncertainty explanation is faithful.

2.4 Benchmark on Tabular Data

Synthetic Data Generation

Evaluating explainability methods on real-world data is challenging due to the subjective nature of interpreting explanations based on expert prior knowledge. To address this, we employ synthetic data with a known data-generating process. Thereby, we can introduce controlled sources of heteroscedasticity, which we aim to detect. Specifically, we sample a synthetic ground truth using a linear system $\boldsymbol{\mu} = \boldsymbol{V}\boldsymbol{\beta}$ with a design matrix $\boldsymbol{V} \in \mathbb{R}^{n \times p}$ with $\boldsymbol{V}_{ij} \sim \mathcal{N}(0, 1)$, and ground truth coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\boldsymbol{\beta}_i \overset{\text{iid}}{\sim} \text{Uniform}([-1, 1])$. We introduce heteroscedastic noise sources with an absolute-value transformed polynomial model for the heteroscedastic noise standard deviation: $\boldsymbol{\sigma} = \mid \phi(\boldsymbol{U})\boldsymbol{\gamma} + \boldsymbol{\delta} \mid$, whereby $\boldsymbol{U} \in \mathbb{R}^{n \times p'}$ is a design matrix with $\boldsymbol{U}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$,

$$\phi(u_1, u_2, \ldots, u_{p'}) \to (1, u_1, \ldots, u_{p'}, u_1^2, u_1 u_2, \ldots, u_{p'}^2)$$

is a second degree polynomial feature map, and $\boldsymbol{\delta} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\delta^2 \boldsymbol{I}\right)$ is the uncertainty model error. The ground truth noise coefficients $\boldsymbol{\gamma} \in \mathbb{R}^{\binom{p'+2}{2}}$ have entries sampled from $\boldsymbol{\gamma}_i \sim \text{Uniform}([-1, -0.5] \cup [0.5, 1])$ to avoid negligible effects. We can then sample the target $\boldsymbol{y} \in \mathbb{R}^n$ with

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \alpha \cdot \text{diag}(\boldsymbol{\sigma}^2) + \sigma_\epsilon^2 \boldsymbol{I}\right),$$

where $\alpha \in \mathbb{R}^+$ determines the overall strength of the heteroscedastic uncertainty and $\sigma_\epsilon^2 \in \mathbb{R}^n$ regulates the homoscedastic noise.

For our experiments, we set $\alpha = 2.0$, $\sigma_\epsilon^2 = 0.02$, and $\sigma_\delta^2 = 0.05$ to get non-negligible, feature-dependent noise. We choose $p = 70$ and $p' = 5$ so that the uncertainty sources have to be detected among a larger set of features that do not influence the uncertainty. We sample $n = 41,500$ data points and concatenate both design matrices to attain the input

$\boldsymbol{X}_{(n\times 75)} = \left[\boldsymbol{U}_{(n\times 5)}, \boldsymbol{V}_{(n\times 70)}\right]$ which we split into 32,000 train, 8,000 validation, and 1,500 test instances.

In reality, we expect noise features to overlap with features influencing the mean. We separate these in the synthetic data to allow for unambiguous assessments in the evaluation. However, our implementation accommodates the analysis of mixed scenarios, where a subset of features simultaneously influences the mean and variance.

Tabular Real World Datasets

In addition, we incorporate three standard regression benchmark datasets into our evaluation: UCI Wine Quality [Cortez *et al.*, 2009], Ailerons [Torgo, 1999], and LSAT academic performance [Wightman, 1998]. These datasets were selected to vary in size and complexity. The Wine Quality dataset, where we use red wines only, includes 11 features for 1,599 samples. Ailerons has 40 features and 13,750 samples, while LSAT, the largest dataset, has 21,790 samples with two continuous and two one-hot encoded features. All datasets are split into 70% training, 10% validation, and 20% testing.

Tabular Benchmarking Setup

We divide our tabular benchmark into two stages. First, we qualitatively and quantitatively evaluate the uncertainty explanation methods in a controlled setting on a synthetically generated dataset. Second, we investigate the same methods concerning their local RRA and RMA, faithfulness, and robustness on synthetic and real-world data.

The first stage of our benchmark aims to detect global drivers of uncertainty in a synthetic setting. We fit a deep neural network of four hidden layers with 64, 64, 64, and 32 units and two outputs for the mean and variance prediction. We train using dropout on the first two layers, Adam optimizer and a batch size of 64. We pre-train using the MSE and fine-tune the model using the GNLL as the loss function, selecting weights with the lowest validation loss. We attain a global feature importance measure as the mean absolute variance feature attributions over all or a specific subset of test instances, which we then analyze using GRA and GMA.

We follow the same model training procedure for the second benchmarking stage, evaluating accuracy, faithfulness, and robustness. Estimating the local RRA and RMA for a given method requires prior knowledge of features affecting the explained quantity, i.e., uncertainty. As this is not available for our selected real-world datasets, we augment them with synthetic noise that we aim to detect, effectively creating a semi-synthetic setting. For the three real-world datasets, we consider two scenarios. We add five noise features to the datasets and heteroscedastic Gaussian noise to the targets with a standard deviation correlating with these features. Since the real-world datasets are small, we first use a simple noise model where the absolute sum of the noise features is the standard deviation of the noise distribution, a setting referred to as 1-S. In a second scenario, 50-C, we use the more complex polynomial noise model described in Section 2.4. To provide more data to the model in the complex noise scenario, we replicate each data point in the train sets 50 times before sampling additional uncertainty features and target noise. For the synthetic datasets, we similarly perform experiments with a simple (S) and complex (C) noise model but without adjusting the dataset size.

We evaluate the robustness of the uncertainty explanation methods for each dataset by estimating the local Lipschitz continuity for 200 randomly selected data points from the test set. For each selected point $\boldsymbol{x}_i$, we compute a local Lipschitz estimate $\hat{L}(\boldsymbol{x}_i)$ by introducing 100 perturbations. For each feature, we sample a perturbation from a uniform distribution centered at the feature value with a range of 2% of the range of the feature in the train set (adapted from Wivestad [2023]). This is not applicable to LSAT's categorical features. Instead, we resort to the discrete definition of local Lipschitz continuity. Specifically, we compute $\hat{L}(\boldsymbol{x}_i)$ for 200 data points sampled from the test set such that their neighborhood $\mathcal{N}_\epsilon(\boldsymbol{x}_i)$ with $\epsilon = 0.2$ contains more than five instances.

To evaluate faithfulness, we apply standard Gaussian noise to perturb the three globally most important uncertainty drivers of the test data. We omit the LSAT dataset as it mainly contains categorical features for which continuous perturbations lack meaning.

We note that we only add synthetic noise to estimate accuracy metrics, i.e., when calculating RRA and RMA. For all other experiments, we use the real-world datasets as is.

2.5 Benchmark on Image Data: MNIST+U

To extend our evaluation to a higher-dimensional problem with more realistic feature dependencies, we consider the task of image regression. We introduce the MNIST+U dataset that extends the original MNIST dataset [Deng, 2012] with an uncertainty component. We create 500,000 composite images with labels. For each sample, two 28×28 MNIST digit images are randomly selected and placed into different corners of a 64×64 canvas. The first digit is white and represents the mean (μ_i) of a target Gaussian distribution. The second gray digit represents its standard deviation (σ_i). Thus, we sample the label (y_i) as: $y_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

We split the generated data into train, validation, and test sets consisting of 70%, 10%, and 20% of the data, respectively. For the image benchmark, we apply a CNN with two parallel encoders where one predicts the mean and the other estimates the variance. Each encoder has two convolutional layers (16 and 32 filters), max-pooling, and fully connected hidden layers with dropout and 128, 64, and 32 nodes. We train the model with MSE for 16 epochs, then switch to GNLL loss until the validation loss converges. We use the Adam optimizer and a batch size of 256. We evaluate uncertainty explainers using RMA and RRA by comparing assigned pixel relevance to the ground truth variance and mean masks. To account for explanations extending beyond the masked digits, each mask is dilated by two pixels.

The code for all experiments and to create the MNIST+U dataset is available online on GitHub[1]. We further make the MNIST+U dataset available separately on Zenodo[2].

[1]https://github.com/DILiS-lab/DroPAU

[2]https://doi.org/10.5281/zenodo.15373739

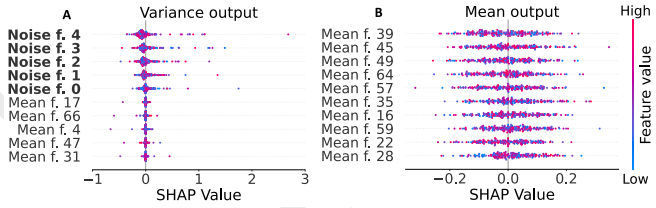

Figure 2: Explanations for uncertainty and mean predictions for the synthetic dataset using VFA-SHAP. We display SHAP summaries for the 10 most important features of (A) model uncertainty or (B) mean prediction ordered by the mean of their absolute estimated Shapley values. VFA-SHAP identifies all noise features driving the model's aleatoric uncertainty. Explaining the mean output offers complementary information but does not detect uncertainty features.

3 Results

3.1 Benchmarking the Detection of Uncertainty Drivers Using Synthetic Datasets

We first examine the capability of VFA-SHAP to identify the drivers of uncertainty, which are features that correlate with the magnitude of the heteroscedastic noise. We know the data-generating process for the synthetic dataset and, therefore, the ground truth noise sources. Using this dataset, VFA-SHAP accurately identifies the five ground-truth noise features driving uncertainty, which are distinct from features influencing the mean (Figure 2 A and B). We verify that our model captures uncertainty accurately to ensure meaningful explanations. All trained models are well-calibrated, predictive of model error, and well-suited to our application setting. Details are available in the GitHub repository.

Further, we analyze the global uncertainty explanation abilities of all VFA flavors, CLUE and InfoSHAP (Figure 3). CLUE is applied to the same neural network as VFA, whereas InfoSHAP utilizes XGBoost. Uncertainty estimation may facilitate cautious model application only at high certainty or opting out of model usage due to substantial uncertainty. Therefore, in addition to 200 random instances, we apply the explainers to the test set's 200 highest and lowest uncertainty instances. We find that VFA flavors and CLUE effectively identify uncertainty drivers for high-uncertainty instances, reflected by their GRAs (with five ground truth noise features) close to 1. VFA flavors exhibit superior GMA, which signifies their capacity to disregard irrelevant features. VFA performs reliably for random and low uncertainty examples, while CLUE's performance deteriorates. This suggests that, unlike CLUE, VFA can explain the factors contributing to certainty. InfoSHAP, while underperforming for instances with high uncertainty, clearly outperforms CLUE for random and low uncertainty instances. We provide code for this figure and further examples showing that VFA considerably outperforms CLUE and InfoSHAP in all three settings.

3.2 Local Accuracies, Faithfulness, and Robustness

We evaluate the local RRA and RMA for the real-world datasets and the synthetic dataset in two settings, one simple (1-S, S) and one complex (50-C, C) as described in Section 2.4 (see Table 1). VFA-SHAP outperforms the other explainers over most datasets. Generally, VFA of any flavor performs best, for the simple settings. However, in the complex setting, InfoSHAP consistently outperforms VFA-IG and VFA-LRP and achieves competitive performance to VFA-SHAP on LSAT and Ailerons.

As shown in Figure 3, CLUE assigns similar importance to all features. This effect is also present for InfoSHAP but is less pronounced. They are, therefore, less selective than VFA, potentially causing uncertainty features not to be detected for many instances.

To analyze the robustness, we calculate distributions of local Lipschitz continuity estimates $\hat{L}(\boldsymbol{x}_i)$ over 200 randomly chosen test set instances for each dataset and method (see Figure 4). According to the obtained Lipschitz estimates, VFA-SHAP and VFA-IG are generally more robust than InfoSHAP, CLUE, and VFA-LRP. The methods' individual ranking differs between datasets, suggesting that the choice of the most robust method is subject to the dataset.

When analyzing the faithfulness metric, we find that perturbation of the most important features faithfully reduces the correlation between uncertainties and residuals when

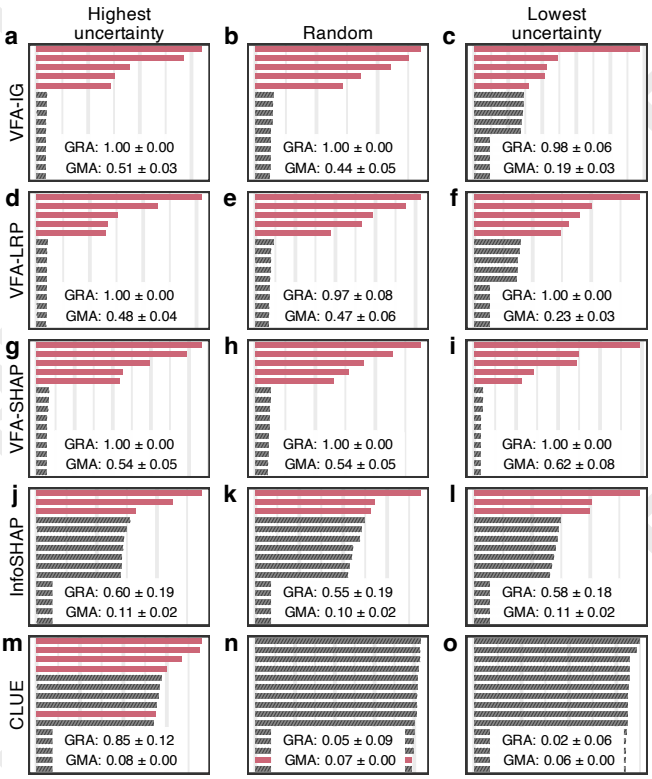

Figure 3: Top 15 global importance features with GRA and GMA for each uncertainty explainer. First column: From 1,500 test samples, we explain the 200 instances with the highest predicted uncertainty. VFA flavors highlight the ground truth noise features (red), while Infoshap and CLUE are less accurate. Second and third columns: For 200 random or low uncertainty instances, VFA remains accurate, while CLUE becomes unreliable. InfoSHAP maintains adequate performance but consistently detects only three noise features.

	Red Wine		Ailerons		LSAT		Synthetic	
	1-S	50-C	1-S	50-C	1-S	50-C	S	C
RRA								
VFA-IG	0.61 ± 0.05	0.60 ± 0.04	0.81 ± 0.03	0.70 ± 0.05	0.81 ± 0.04	0.74 ± 0.05	0.75 ± 0.02	0.38 ± 0.05
VFA-LRP	0.62 ± 0.05	0.61 ± 0.04	0.79 ± 0.04	0.67 ± 0.06	0.81 ± 0.02	0.73 ± 0.06	0.75 ± 0.01	0.41 ± 0.05
VFA-SHAP	$\mathbf{0.85} \pm 0.02$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.88} \pm 0.01$	$\mathbf{0.88} \pm 0.02$	$\mathbf{0.93} \pm 0.02$	$\mathbf{0.92} \pm 0.02$	$\mathbf{0.85} \pm 0.01$	$\mathbf{0.70} \pm 0.06$
CLUE	0.38 ± 0.22	0.65 ± 0.02	0.41 ± 0.12	0.58 ± 0.02	0.54 ± 0.02	0.49 ± 0.01	0.07 ± 0.00	0.07 ± 0.00
InfoShap	0.41 ± 0.06	0.72 ± 0.04	0.52 ± 0.02	$\mathbf{0.88} \pm 0.02$	0.79 ± 0.01	$\mathbf{0.92} \pm 0.01$	0.59 ± 0.02	0.49 ± 0.05
RMA								
VFA-IG	0.57 ± 0.05	0.64 ± 0.03	0.79 ± 0.04	0.72 ± 0.07	0.83 ± 0.04	0.81 ± 0.08	0.50 ± 0.02	0.25 ± 0.03
VFA-LRP	0.57 ± 0.05	0.65 ± 0.04	0.76 ± 0.04	0.70 ± 0.08	0.82 ± 0.04	0.86 ± 0.05	0.49 ± 0.01	0.26 ± 0.03
VFA-SHAP	$\mathbf{0.83} \pm 0.03$	$\mathbf{0.92} \pm 0.01$	$\mathbf{0.89} \pm 0.02$	$\mathbf{0.87} \pm 0.05$	$\mathbf{0.95} \pm 0.03$	$\mathbf{0.94} \pm 0.02$	$\mathbf{0.75} \pm 0.02$	$\mathbf{0.44} \pm 0.09$
CLUE	0.34 ± 0.17	0.60 ± 0.02	0.27 ± 0.11	0.47 ± 0.01	0.52 ± 0.02	0.50 ± 0.01	0.07 ± 0.00	0.07 ± 0.00
InfoShap	0.38 ± 0.04	0.67 ± 0.04	0.41 ± 0.01	0.83 ± 0.03	0.79 ± 0.02	$\mathbf{0.94} \pm 0.01$	0.31 ± 0.01	0.26 ± 0.03

Table 1: Average local RRA and RMA over all test set instances for all considered uncertainty explainers and datasets (1-S: simple noise model and original train set, 50-C: complex noise model and artificially enlarged train set). Results are averaged across five folds, with standard deviations shown in brackets and best performances in bold. VFA consistently outperforms InfoSHAP and CLUE in all simple scenarios. VFA-SHAP also consistently performs best in all complex scenarios, while VFA-LRP and VFA-IG outperform CLUE in most cases and InfoSHAP in some complex scenarios.

VFA is used in the Ailerons and synthetic datasets (see Table 2). However, VFA-LRP exhibits weaker faithfulness on the Ailerons dataset, demonstrating performance comparable to the baseline methods.

On the small Red Wine dataset, the faithfulness of all methods near zero. In essence, the challenge of learning and explaining uncertainty is amplified in scenarios where data are scarce, leading to suboptimal faithfulness metrics.

3.3 Benchmark on MNIST+U Image Data

We evaluate the variance explainers on the MNIST+U dataset to understand their capability of dealing with higher-dimensional image data. We expect high attributions for pixels in the area of the uncertainty mask and, relative to that, low attribution on the mean mask.

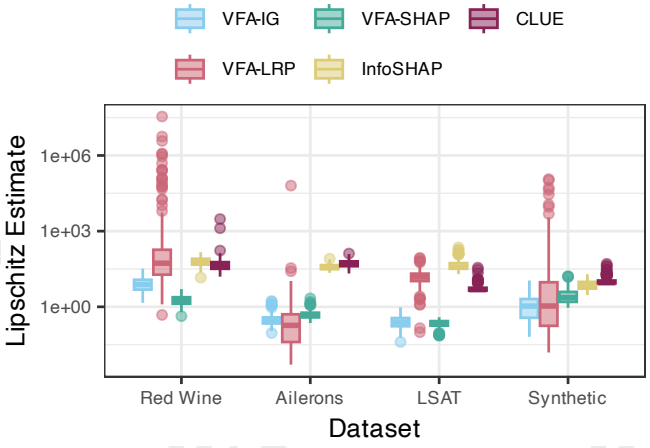

Figure 4: Local Lipschitz continuity estimates for 200 randomly chosen test set instances for all methods and datasets. Lower values indicate higher robustness. Having the lowest median Lipschitz estimates for most datasets, VFA-SHAP and VFA-IG are the generally more robust explainers.

All explanation methods, excluding VFA-IG, focus on the area of uncertainty mask (see Figure 5). VFA-LRP performs best, demonstrating the highest relevance attribution to the variance mask. This observation aligns with our expectation that most relevance should correspond to the variance, representing the primary source of uncertainty in the synthetic labels. We also see this in a randomly selected explanation example for VFA-LRP shown in Figure 6.

While InfoSHAP and CLUE assign a considerable amount of relevance to the variance mask, they also have larger proportions of the variance explanation assigned to the mean mask.

VFA-IG performs poorly and assigns similar amounts of relevance to mean and variance masks, highlighting that the choice of the explanation method strongly influences the results. This aligns with findings that IG explanations on images focus on specific pixels rather than relevant patterns driving the prediction [Samek *et al.*, 2021]. This emphasizes the importance of selecting an appropriate explainer tai-

	Red Wine	Ailerons	Synthetic
VFA-IG	-0.001 ± 0.017	-0.139 ± 0.064	$\mathbf{-0.167} \pm 0.026$
VFA-LRP	-0.004 ± 0.020	-0.083 ± 0.042	-0.154 ± 0.028
VFA-SHAP	0.003 ± 0.018	$\mathbf{-0.171} \pm 0.026$	-0.158 ± 0.032
InfoSHAP	-0.001 ± 0.054	-0.099 ± 0.036	-0.082 ± 0.023
CLUE	-0.000 ± 0.016	-0.025 ± 0.009	-0.009 ± 0.015

Table 2: Faithfulness of the uncertainty explanations: Change of Spearman correlation between uncertainties and squared residuals when most important features are perturbed. We expect faithful uncertainty explanations to induce a negative change. We exclude LSAT because we define perturbations only for continuous features. Results are the mean and standard deviation of 12 folds.

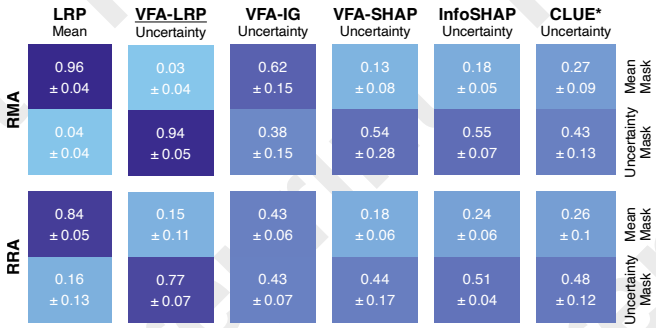

Figure 5: RMA and RRA for each uncertainty explainer and LRP mean explanations. We compare the attribution of pixels in the ground truth mask of the mean or the noise. We expect most of the relevance to be contained in the uncertainty mask. We show the mean and standard deviation over all samples in the test set.
(*) Note: for CLUE, we only use 40% of the test samples due to its high runtime.

lored to the specific characteristics of the data and uncertainty sources, which can be achieved in practice by considering faithfulness on a validation set.

4 Discussion and Limitations

We presented a straightforward strategy for explaining predictive aleatoric uncertainties, which requires minimal modifications to existing neural network regressors. We use neural networks with a Gaussian output distribution to estimate uncertainty and apply explanation methods to the variance output to explain uncertainty. In synthetic tabular experiments, the resulting explanations generally outperform alternative methods. As seen in the experiments with low uncertainty instances, we can also explain how features contribute to a model's certainty, which is relevant in high-risk applications. Since conventional evaluation metrics are not always directly applicable, we have introduced an evaluation protocol to assess uncertainty explainers. Parts of our evaluation depend on the knowledge of ground truth noise sources. This necessitated the incorporation of synthetic noise, which may deviate from the arbitrarily complex real-world noise patterns. We extend unsupervised explanation quality metrics for accuracy, faithfulness, and robustness to uncertainty attributions. In our benchmark, VFA compares favorably to CLUE and InfoSHAP. Using the MNIST+U image benchmark, we establish

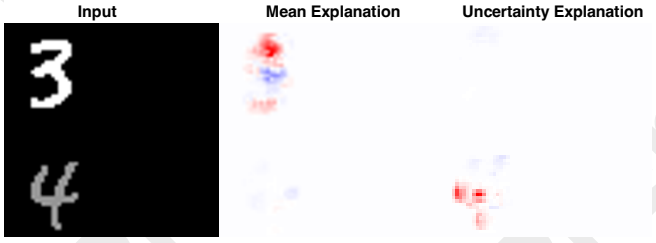

Figure 6: Mean and uncertainty explanations using LRP/VFA-LRP for a random sample from the MNIST+U test set. Both explanations focus on the digits relevant for mean and uncertainty, respectively.

that the selection of the explanation method is a significant variable in generating high-fidelity uncertainty explanations. Generally, as we combine deep heteroscedastic regression with existing XAI methods, we inherit all the benefits and limitations of these methods, including computational complexity. A limitation of our analysis is its focus on aleatoric uncertainty. Approaches to model and explain epistemic uncertainty are equally crucial for safety-critical settings and should be used alongside our approach [Bley *et al.*, 2025]. Future work might involve studying synergies in explaining point and uncertainty predictions. For example, in the context of explainable active learning [Ghai *et al.*, 2021], a visualization of both explainability modes could be beneficial.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work is supported by a BMWK grant (DAKI, 01MK21009E), a BMBF grant (act-i-ml, 01IS24078B), and a European Research Council grant (eXplAInProt, 101124385).

Contribution Statement

PI and SW contributed equally to this work. KW and BYR share senior authorship.

References

[Abdar *et al.*, 2021] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, 76:243–297, December 2021. arXiv:2011.06225 [cs].

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[Almosallam *et al.*, 2016] Ibrahim A. Almosallam, Matt J. Jarvis, and Stephen J. Roberts. GPz: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739, July 2016.

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S. Jaakkola. On the Robustness of Interpretability Methods. In *WHI 2018*, June 2018.

[Antoran *et al.*, 2021] Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A Method for Explaining Uncertainty Estimates. In *International Conference on Learning Representations (ICLR)*, 2021.

[Arras *et al.*, 2022] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. Publisher: Elsevier.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, October 2015.

[Bauza and Rodriguez, 2017] Maria Bauza and Alberto Rodriguez. A probabilistic data-driven model for planar pushing. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3008–3015, 2017.

[Bhatt *et al.*, 2021] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 401–413, New York, NY, USA, July 2021. Association for Computing Machinery.

[Bishop, 1994] Christopher Bishop. Mixture density networks. *Aston University Neural Computing Research Group Report*, 1994.

[Bley *et al.*, 2025] Florian Bley, Sebastian Lapuschkin, Wojciech Samek, and Grégoire Montavon. Explaining predictive uncertainty by exposing second-order effects. *Pattern Recognition*, 160:111171, April 2025.

[Brown and Talbert, 2022] Katherine Elizabeth Brown and Douglas A. Talbert. Using Explainable AI to Measure Feature Contribution to Uncertainty. *The International FLAIRS Conference Proceedings*, 35, May 2022.

[Chua *et al.*, 2023] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G. Lee, Vikram Deshpande, Joseph Schwab, Michael H. Lev, Ramon G. Gonzalez, Michael S. Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, June 2023.

[Cortez *et al.*, 2009] Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality, 2009. Published: UCI Machine Learning Repository.

[Deng, 2012] Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142, November 2012. Conference Name: IEEE Signal Processing Magazine.

[Ghai *et al.*, 2021] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):235:1–235:28, January 2021.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5580–5590, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Kompa *et al.*, 2021] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6, January 2021.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Ley *et al.*, 2022] Dan Ley, Umang Bhatt, and Adrian Weller. Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7390–7398, June 2022.

[Liu *et al.*, 2021] Haitao Liu, Yew-Soon Ong, and Jianfei Cai. Large-Scale Heteroscedastic Regression via Gaussian Process. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):708–721, 2021.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Lázaro-Gredilla *et al.*, 2014] Miguel Lázaro-Gredilla, Michalis K. Titsias, Jochem Verrelst, and Gustavo Camps-Valls. Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes. *IEEE Geoscience and Remote Sensing Letters*, 11(4):838–842, 2014.

[Lötsch *et al.*, 2022] Jörn Lötsch, Dario Kringel, and Alfred Ultsch. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics*, 2(1):1–17, March 2022.

[MacKay, 1992] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, May 1992.

[McGrath *et al.*, 2023] Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. *Transactions on Machine Learning Research*, February 2023.

[Mehdiyev *et al.*, 2023] Nijat Mehdiyev, Maxim Majlatow, and Peter Fettke. Quantifying and Explaining Machine Learning Uncertainty in Predictive Process Monitoring: An Operations Research Perspective, April 2023. arXiv:2304.06412 [cs, math, stat].

[Mougan and Nielsen, 2023] Carlos Mougan and Dan Saattrup Nielsen. Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric

Bootstrap. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15037–15045, June 2023.

[Nix and Weigend, 1994] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, June 1994.

[Perez *et al.*, 2022] Iker Perez, Piotr Skalski, Alec Barns-Graham, Jason Wong, and David Sutton. Attribution of predictive uncertainties in classification models. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 1582–1591. PMLR, August 2022.

[Samek *et al.*, 2021] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3):247–278, March 2021.

[Schwalbe and Finzel, 2023] Gesina Schwalbe and Bettina Finzel. A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts. *Data Mining and Knowledge Discovery*, January 2023. arXiv:2105.07190 [cs].

[Seitzer *et al.*, 2022] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks, April 2022. arXiv:2203.09168 [cs, stat].

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3145–3153, Sydney, NSW, Australia, August 2017. JMLR.org.

[Sluijterman *et al.*, 2023] Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal Training of Mean Variance Estimation Neural Networks, August 2023. arXiv:2302.08875 [cs, stat].

[Smith *et al.*, 2018] Andrew J. Smith, Mohammed Al-Absi, and Travis Fields. Heteroscedastic Gaussian Process-based System Identification and Predictive Control of a Quadcopter. In *2018 AIAA Atmospheric Flight Mechanics Conference*. American Institute of Aeronautics and Astronautics, July 2018. _eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2018-0298.

[Sundararajan *et al.*, 2017] 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, pages 3319–3328, Sydney, NSW, Australia, August 2017. JMLR.org.

[Torgo, 1999] Luís Fernando Raínho Alves Torgo. *Inductive learning of tree-based regression models*. Ph.D. thesis, University of Porto, 1999.

[Vilone and Longo, 2020] Giulia Vilone and Luca Longo. Explainable Artificial Intelligence: a Systematic Review, October 2020. arXiv:2006.00093 [cs].

[Wang *et al.*, 2023] Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, and Qiang Ji. Gradient-Based Uncertainty Attribution for Explainable Bayesian Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12044–12053, 2023.

[Watson *et al.*, 2023] David S. Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining Predictive Uncertainty with Information Theoretic Shapley Values, June 2023. arXiv:2306.05724 [cs, stat].

[Wightman, 1998] Linda F Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998.

[Wivestad, 2023] Viggo Tellefsen Wivestad. https://github.com/viggotw/Robustness-of-Interpretability-Methods, May 2023. original-date: 2021-12-03T18:24:06Z.

[Wong-Toi *et al.*, 2023] Eliot Wong-Toi, Alex Boyd, Vincent Fortuin, and Stephan Mandt. Understanding Pathologies of Deep Heteroskedastic Regression, June 2023. arXiv:2306.16717 [cs, stat].

[Yang and Li, 2023] Chu-I Yang and Yi-Pei Li. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15(1):13, February 2023.