Explaining Set-Valued Predictions: SHAP Analysis for Conformal Classification

Ulf Johansson ULF.JOHANSSON@JU.SE
Dept. of Computing, Jönköping University, Sweden

Aicha Maalej AICHA.MAALEJ@JU.SE
Dept. of Computing, Jönköping University, Sweden
School of Informatics, University of Skövde, Sweden

Cecilia Sönströd CECILIA.SONSTROD@JU.SE
Dept. of Computing, Jönköping University, Sweden

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Conformal prediction offers a principled framework for uncertainty quantification in classification tasks by outputting prediction sets with guaranteed error control. However, the interpretability of these set-valued predictions, and consequently their practical usefulness, remains underexplored. In this paper, we introduce a method for explaining conformal classification outputs using SHAP (SHapley Additive exPlanations), enabling model-agnostic local and global feature attributions for the p-values associated with individual class labels. This approach allows for rich, class-specific explanations in which feature effects need not be symmetrically distributed across classes. The resulting flexibility supports the detection of ambiguous predictions and potential out-of-distribution instances in a transparent and structured way. While our primary focus is on explaining p-values, we also outline how the same framework can be applied to related targets, including label inclusion, set predictions, and the derived confidence and credibility measures. We demonstrate the method on several benchmark datasets and show that SHAP-enhanced conformal predictors offer improved interpretability by revealing the drivers behind set predictions, thereby providing actionable insights in high-stakes decision-making contexts.

Keywords: Conformal classification, Set predictions, XAI, Local explanations, Global explanations, SHAP

1. Introduction.

Data-driven decision making is increasingly prevalent across domains such as healthcare, drug discovery, finance, retail, and manufacturing. Predictive models created by machine learning are now routinely used to support or automate decisions. However, for such systems to be trusted and integrated into real-world workflows, they must not only be accurate but also provide accurate uncertainty quantification.

Instead of single class labels or estimated probability distributions, conformal classifiers output a set of class labels that is guaranteed, under mild assumptions, to contain the true label with a predefined level of confidence. Conformal prediction thus provides a theoretically sound way to control errors and reason about uncertainty in classification.

However, despite its attractive theoretical properties, the interpretability and the usefulness of conformal predictions remain relatively underexplored. In particular, when prediction sets are non-trivial (e.g., contain multiple labels or no lables), it may be difficult for users to understand why the model is uncertain, or why specific labels are included or excluded. Traditional feature attribution methods, such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017),

have been widely adopted for interpreting probabilistic models, but have not yet been extended in a straightforward way to the set predictions from conformal predictors.

In this paper, we propose a method for explaining prediction sets produced by conformal classifiers using SHAP. We focus primarily on the p-values generated for each class by the conformal framework, treating them as model outputs suitable for local and global SHAP analysis. This enables feature attribution for each label's p-value and offers a fine-grained view of how input features contribute to uncertainty, confidence, and ambiguity in the predictions. Importantly, because conformal p-values are computed independently for each class, SHAP values can reveal asymmetric or class-specific feature effects. This offers a key advantage over standard explanations of probabilistic predictions, where class probabilities are constrained to sum to one, potentially masking such effects.

While our main contribution is the SHAP-based explanation of p-values, we also outline and evaluate additional ways of applying SHAP to conformal classifiers. These include explaining the inclusion or exclusion of individual labels, the predicted label set as a whole, as well as the derived confidence and credibility measures. Together, these alternatives demonstrate the flexibility of SHAP for explaining different properties of conformal classification outputs.

We demonstrate our approach on a number of benchmark datasets, illustrating its practical value in supporting decision making. Specifically, we argue that SHAP-enhanced conformal prediction improves both the interpretability and trustworthiness of model outputs, especially in contexts where reliable uncertainty quantification is essential.

2. Background.

2.1. Explanations

Interpretability has become a central concern in the development and deployment of machine learning models, see e.g., (Molnar et al., 2020). When model predictions influence real-world outcomes, it is essential that end-users, whether domain experts, affected individuals, or regulators, can gain insight not only into the model's outputs but also into the underlying reasons for its predictions. This transparency supports trust, accountability, and informed decision-making, and is increasingly seen as a prerequisite for ethical and responsible AI.

One approach to achieving interpretability is through the use of inherently interpretable models, such as decision trees, rule-based systems, or linear models, which provide explanations directly through their structure. While these models offer valuable transparency, they may not always achieve the predictive performance of more complex, opaque models like ensemble methods or (deep) neural networks. To reconcile the need for both performance and interpretability, a complementary strategy has emerged: post-hoc explanation methods that are applied after a model is trained, treating it as a black box.

Explanations generated post-hoc come in several forms. Some aim to provide global explanations, describing the overall behavior of the model across the dataset, while others focus on local explanations, which account for why the model made a specific prediction for a given input. Another strategy is based on counterfactual explanations, which indicate how an input would need to change in order to yield a different prediction.

Feature attribution methods aim to provide insight into how individual features contribute to a model's prediction. Specifically, Permutation Feature Importance is a family of methods, based on sensitivity analysis. These methods are widely used for model interpretation, and are available in many machine learning libraries, including scikit-learn. The key idea is to randomly permute a

feature's values and measure the change in model performance, i.e., a large decrease in performance indicates that the feature is highly important for the model's prediction.

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) explains a model's prediction by approximating it locally with an interpretable, simpler model (such as a linear regression). It perturbs the input data and observes how the model's output changes, providing feature importance values based on this local approximation.

SHAP (Lundberg and Lee, 2017) is a widely used framework for interpreting the predictions of machine learning models. Rooted in cooperative game theory, SHAP assigns each feature a contribution value that reflects its importance to a specific prediction. These contribution values, called SHAP values, are grounded in the concept of Shapley values, originally developed to fairly distribute gains among players in a cooperative game. In the context of explaining machine learning predictions, each "player" is a feature, and the "game" is the prediction made by the model. SHAP directly answers how much a feature value influenced the model's output for the particular instance. The key idea is to consider all possible combinations of features and evaluate how adding each one affects the prediction, averaging this effect across all such combinations.

2.2. Conformal classification

In conformal classification, test instances are tentatively labeled as $(\boldsymbol{x}_{k+1}, \tilde{y})$, and a p-value statistic is then calculated to assess whether the hypothesis that $\tilde{y} = y_{k+1}$ can be rejected at a given significance level ϵ. This procedure is repeated for all possible labels, resulting in a predicted label set $\tilde{y} \subseteq Y$ that includes all labels not rejected. Under mild assumptions, this label set is guaranteed to contain the true label y_{k+1} with probability of at least $1 - \epsilon$.

To determine which labels can be rejected, a so-called *nonconformity function* $A : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is used. This function quantifies how unusual a labeled instance $(\boldsymbol{x}, \tilde{y})$ is, relative to a set of calibration examples with known labels. If the p-value for a given label is below the threshold ϵ, the label is rejected.

In predictive modeling, the nonconformity function is typically based on the prediction scores of a machine learning model, referred to as the *underlying model*. In this study, we use the *hinge loss* function,

$$\Delta\left[h\left(\boldsymbol{x}_i\right), \tilde{y}\right] = 1 - \hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}_i\right), \tag{1}$$

where $\hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}_i\right)$ denotes the model's predicted probability for label $\tilde{y}$ given input $\boldsymbol{x}_i$.

An *inductive conformal predictor* (ICP) (Papadopoulos et al., 2002; Vovk et al., 2005; Papadopoulos, 2008) for classification is constructed in the following way:

1. Split the labeled dataset Z into two disjoint subsets: a proper training set Z^t and a calibration set Z^c, where $|Z^c| = q$.

2. Train the underlying model h using the proper training set Z^t.

3. Apply the nonconformity function (Eq. 1) to all calibration examples in Z^c to obtain a list of nonconformity scores $\alpha_1, \ldots, \alpha_q$.

To obtain the predicted label set for a test instance $\boldsymbol{x}_{k+1}$:

1. Use the trained model to obtain the prediction $h(\boldsymbol{x}_{k+1})$.

2. For each possible label $\tilde{y} \in Y$, compute the nonconformity score $\alpha_{k+1}^{\tilde{y}}$ for the test pair $(\boldsymbol{x}_{k+1}, \tilde{y})$.

3. Compute the corresponding p-value as:

$$p_{k+1}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z_c : \alpha_i > \alpha_{k+1}^{\tilde{y}}\right\}\right|}{q+1} + \theta_{k+1}\frac{\left|\left\{z_i \in Z_c : \alpha_i = \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{q+1}, \tag{2}$$

   where $\theta_{k+1} \sim \mathcal{U}[0,1]$.

4. Reject all labels $\tilde{y}$ for which $p_{k+1}^{\tilde{y}} < \epsilon$.

5. The final prediction set Γ_{k+1}^{ϵ} consists of all labels not rejected.

This procedure guarantees that the probability of the prediction set Γ_{k+1}^{ϵ} containing the true label is at least $1-\epsilon$, assuming that the calibration and test instances are exchangeable.

By conditioning the p-value computation on specific attributes of the test instance, one obtains a so-called *Mondrian* conformal classifier (Vovk et al., 2005), which allows for category-wise error guarantees. The most common case is the class-conditional conformal classifier, where p-values are computed using only calibration examples that share the same label. In this case, Eq. (2) is replaced by:

$$p_{k+1}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z^{\tilde{y}} : \alpha_i > \alpha_{k+1}^{\tilde{y}}\right\}\right|}{|Z^{\tilde{y}}|+1} + \theta_{k+1}\frac{\left|\left\{z_i \in Z^{\tilde{y}} : \alpha_i = \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{|Z^{\tilde{y}}|+1}, \tag{3}$$

where $Z^{\tilde{y}} \subseteq Z_c$ is the subset of calibration examples with true label $\tilde{y}$.

2.2.1. Confidence and Credibility

Given a conformal predictor that outputs a set of p-values $\{p_y(\boldsymbol{x}_j)\}_{y\in\mathcal{Y}}$ for an instance $\boldsymbol{x}_j$, for each class label $y \in \mathcal{Y}$, the following quantities can be defined:

- **Credibility** is the largest p-value across all classes:

$$\text{Credibility}(x) = \max_{y\in\mathcal{Y}} p_y(x) \tag{4}$$

  Credibility reflects how well the most plausible class label conforms to the model's learned distribution.

- **Confidence** is one minus the second-largest p-value:

$$\text{Confidence}(x) = 1 - \max_{y\in\mathcal{Y}\setminus\{y^*\}} p_y(x) \tag{5}$$

  where

$$y^* = \arg\max_{y\in\mathcal{Y}} p_y(x) \tag{6}$$

  Thus, confidence measures the plausibility of the second most likely label, i.e., when all labels except one can be rejected.

Loosely put, confidence measures how decisive the prediction is, i.e., it reflects how strongly the most plausible class stands out compared to the others, while credibility measures how well the instance conforms to any known class. In other words, credibility captures how typical or strange the instance is with respect to the calibration data.

The connection between confidence-credibility measures and the set predictions is that confidence represents the highest significance level where we get a singleton prediction, while the credibility is the highest significance level where all labels are rejected.

2.2.2. Conformal classifiers for decision support

Conformal classifiers support decision-making by producing prediction sets that control the risk of misclassification at a user-defined significance level. Rather than forcing a single label prediction, the model returns a set of plausible labels, each accompanied by a p-value that reflects how well the observed data conforms to that label. This allows decision-makers to see not only what the model predicts, but also how uncertain it is about that prediction.

In binary classification, conformal prediction provides three possible outputs: a singleton set containing either class, i.e., $\{0\}$ or $\{1\}$, a full set containing both classes $\{0, 1\}$, or, in rare cases, an empty set $\{\}$, when no label meets the required significance threshold. A singleton prediction implies high confidence, while a full set reflects higher model uncertainty or model ambiguity. An empty prediction set signals that the model cannot support either class at the given confidence level, this can be interpreted as a strong sign of out-of-distribution input, extreme uncertainty, or possibly data corruption.

In multiclass problems, conformal classification reveals even more utility. Rather than outputting a single label, the predictor returns a subset of plausible classes based on a predefined significance level. Importantly, it can explicitly reject unlikely classes, offering safe zones of reasoning where low-probability decisions are not forced on the user.

This kind of output becomes particularly meaningful in real-world scenarios with ambiguous or overlapping classes. For example, in industrial quality control or disease diagnosis, it may be more helpful to rule out certain causes than to attempt overly confident predictions. Prediction sets provide a compact way to reflect epistemic uncertainty, especially when a definitive answer cannot be given.

The associated confidence and credibility scores produced by the conformal predictor provide further insight into the reliability of the prediction set. An instance with high confidence but low credibility might suggest a "least-worst" prediction where the model picked something, but didn't really trust it. An instance with low confidence but high credibility might mean that multiple classes look equally plausible, i.e., the case is ambiguous. Finally, low credibility suggests out-of-distribution behavior or at least a rare/unseen pattern.

For instance, in a medical setting, a singleton prediction with high confidence might support a clear diagnosis, while a multi-label set with low confidence may indicate the need for further testing or expert review. In contrast, standard classifiers provide only the most likely label, and probabilistic predictors output a normalized distribution over classes that can sometimes overstate certainty, particularly in the presence of distribution shift or limited data.

The ability to communicate predictions like "we're not sure, but it's likely one of these few options, and definitely not this one" adds practical value that is often lost in standard probabilistic predictors, which are constrained to give normalized probabilities, even when none of the classes are truly plausible.

2.3. SHAP: SHapley Additive exPlanations

SHAP is a method for explaining the output of machine learning models by assigning each input feature a contribution value, known as a SHAP value, for a given prediction.

SHAP explains a model's prediction $f(x)$ for an instance $x \in \mathbb{R}^d$ by computing feature contributions ϕ_i such that:

$$f(x) = \phi_0 + \sum_{i=1}^{d} \phi_i \tag{7}$$

Here:

- ϕ_0 is the expected model output over the training data, i.e., $\phi_0 = \mathbb{E}[f(x)]$.
- ϕ_i is the Shapley value corresponding to feature i, representing its marginal contribution to the prediction.

Shapley values originate from cooperative game theory and provide a way to fairly distribute the total "payout" among players (features), assuming they work together in all possible coalitions. For feature i, the Shapley value is defined as:

$$\phi_i(f, x) = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f_{S \cup \{i\}}(x) - f_S(x) \right] \tag{8}$$

where:

- S is a subset of features not containing i,
- $f_S(x)$ is the model output when only features in S are known,
- The weight is a combinatorial term ensuring fairness.

SHAP values uniquely satisfy the following properties, which make them desirable for model explanations:

1. **Local Accuracy (Additivity):** The explanation model matches the original model output for any instance.
2. **Missingness:** If a feature is missing (or has no impact), its attribution is zero.
3. **Consistency:** If a model changes such that the marginal contribution of a feature increases, its SHAP value does not decrease.

Exact computation of Shapley values is exponential in the number of features, but SHAP provides several approximations:

- `KernelExplainer` for model-agnostic approximation using weighted linear regression.
- `TreeExplainer` (Lundberg et al., 2020) for efficient exact computation on tree-based models.
- `DeepExplainer` and `GradientExplainer` for neural networks.

SHAP values allow both **local explanations** (per-instance feature attributions) and **global insights** (aggregated feature importance across a dataset). There are also several plots available. In this paper, we will use *Waterfall plots* for local additive decompositions and *Beeswarm plots* for global distributions of SHAP values. In summary, the SHAP framework has several appealing properties:

- Model-agnostic and model-specific options: SHAP can be applied to any model via sampling-based methods (like KernelExplainer), and also has efficient implementations for tree-based models (like random forests, XGBoost and LightGBM).
- Local explanations: SHAP explains individual predictions, showing how each feature value pushed the output higher or lower.

- Global insights: By aggregating SHAP values over many instances, one can get a picture of overall feature importance and model behavior.
- Additivity: The explanation satisfies the property that the sum of all feature contributions equals the model's output (relative to a baseline), supporting intuitive, faithful interpretation.
- Visualizations: Many different and complementing plots readily available.

2.4. Related Work

Conformal prediction (CP) has emerged as a principled framework for providing valid uncertainty quantification in machine learning, particularly valuable for decision-making under uncertainty. One of the earliest and most impactful application areas has been in drug discovery and medicinal chemistry, where CP has contributed to the development of reliable, uncertainty-aware predictive models. Much early work was conducted within the domain of Quantitative Structure–Activity Relationship (QSAR) modeling. In particular, Eklund et al. (2012) introduced CP into QSAR workflows to produce prediction intervals and improve model transparency. This was further developed by Norinder et al. (2014), who emphasized CP's potential to support regulatory decision-making by offering class-specific confidence measures. Building on this foundation, Sun et al. (2017) applied Mondrian Cross-Conformal Prediction (MCCP) to large, imbalanced bioactivity datasets, highlighting CP's robustness and reliability even in challenging predictive scenarios.

These early efforts demonstrated how CP could effectively quantify predictive uncertainty at the individual compound level, an essential component in decision-making pipelines in cheminformatics. Cortés-Ciriano and Bender (2019) further emphasized CP's utility in virtual screening and bioactivity modeling, citing its minimal computational cost and seamless integration with existing pipelines. Similarly, Ahlberg et al. (2017) showed how CP could guide synthesis prioritization by quantifying uncertainty and improving the interpretability of compound rankings.

Several recent studies have expanded the application of CP to modern machine learning settings. Ghosh et al. (2023) proposed Neighborhood Conformal Prediction (NCP), an efficient CP technique tailored to deep neural networks, demonstrating improved uncertainty quantification on multiple benchmarks. Svensson et al. (2019) explored conformal regression for producing calibrated prediction intervals, while Krstajic (2020) provided a large-scale comparison of CP and QSAR methods, reinforcing the utility of CP for interpretable and confident decision support.

Integrating conformal prediction with interpretable classification models, such as decision trees and rule-based classifiers, has been explored to combine the benefits of model transparency and reliable uncertainty quantification. This is particularly valuable in domains where understanding the rationale behind predictions is as crucial as the predictions themselves. Johansson et al. (2013b,a) investigated the application of CP to decision trees, demonstrating that CP can provide valid prediction sets while maintaining the interpretability of tree-based models. Hüllermeier et al. (2020) explored the use of CP in multi-label classification using rule-based models. They demonstrated that rules can naturally provide non-conformity scores required by CP, and that CP can calibrate the assessment of candidate rules, thereby supporting better predictions and more elaborate decision-making. Abdelqader et al. (2023) proposed *conformal decision rules*, which combine the transparency of rule-based classifiers with the reliability of CP. Their technique provides point predictions accompanied by confidence measures, facilitating trustworthy and understandable model outputs. Tyagi and Guo (2024) introduced a tree-based conformal prediction approach for multi-label classification under uncertainty. Their method employs hierarchical clustering with label sets to develop a hierarchical tree structure, formulating the problem as a multiple-testing scenario with a hierarchical structure, and providing theoretical guarantees for valid coverage.

Other recent advancements in explainable AI have emphasized the importance of not only providing model explanations but also quantifying the confidence associated with these explanations. Löfström et al. (2024) introduced *Calibrated Explanations*, a method that offers feature importance explanations accompanied by uncertainty estimates, utilizing Venn-Abers predictors to ensure well-calibrated outputs. This approach has been extended to conformal regression and conformal predictive systems (Löfström et al., 2025).

Alkhatib et al. (2022) proposed the use of Venn prediction to assess the quality of explanations, enabling the quantification of uncertainty in rule-based explanations and facilitating the evaluation of different explanation techniques based on their predictive reliability. Further, they explored approximating complex explanation techniques like SHAP using conformal regression, providing validity guarantees for the approximated explanations (Alkhatib et al., 2023).

While some recent studies have combined conformal prediction with feature attribution or explanation techniques, these efforts have primarily focused on using conformal methods to quantify uncertainty in explanations themselves. To the best of our knowledge, this is the first work that explicitly aims to explain the conformal predictions, that is, to provide feature-level attributions for why a particular prediction set was returned. Specifically, we propose the use of SHAP values to interpret set-valued outputs from conformal predictors, offering insight into how individual features contribute to the inclusion or exclusion of specific labels in the prediction set.

3. Method

3.1. SHAP analysis for conformal classifiers

In this paper, we extend the use of SHAP beyond traditional model outputs, such as predicted probabilities or class scores, to explain the p-values produced by conformal classification. This opens up new possibilities for understanding not just what the model predicts, but also how confident it is in making those predictions, and why.

When SHAP is applied to the p-values output by conformal classifiers in binary classification, it produces class-specific local explanations for each instance. This means that users can understand why one class received a high p-value (i.e., why it was included in the prediction set) and why another did not. Crucially, these explanations are not mirror images of each other, unlike in standard probabilistic models where class probabilities must sum to one.

This flexibility opens up for richer and more nuanced interpretations. For instance, in a binary medical diagnosis, SHAP can explain the inclusion of a disease label based on elevated lab values, while the exclusion of the other label might stem from a different subset of features (e.g., absence of risk factors). This type of asymmetric insight allows practitioners to weigh factors independently for each hypothesis, rather than relying on relative probabilities alone.

In rare but important cases, conformal predictors may output an empty prediction set, signaling that neither class met the required significance level. Rather than treating this as a failure or black-box anomaly, SHAP can offer transparency by explaining why both classes were excluded, highlighting which features drove down the p-values. This can be particularly valuable for detecting out-of-distribution instances or borderline cases where no confident prediction can be made. In such situations, SHAP explanations can help to guide appropriate further action, such as requesting additional information or expert review.

In multiclass settings, the combination of conformal prediction and SHAP can be used to explain why certain classes were rejected, which is a use case that probabilistic classifiers typically do not support in a transparent way. Each p-value gets its own SHAP decomposition, enabling an understanding of which features contributed to excluding specific labels, and which supported

inclusion. In many decision support systems, this can provide important insights. For example, in a clinical triage tool considering multiple diagnoses, it may be critical to understand why a dangerous condition was not predicted. SHAP allows the model to say, "this disease was excluded because features A, B, and C were not present," even if other diseases remain plausible. Additionally, across a population, global SHAP analyses (e.g., beeswarm or summary plots) can identify consistent patterns behind prediction set sizes and composition.

While the primary focus of this work is on explaining p-values, the same SHAP-based framework can be applied to other conformal outputs. These include explaining the inclusion or exclusion of individual labels, the full predicted label set, as well as the derived confidence and credibility scores.

3.2. Explaining Conformal Predictions with SHAP

In practice, the outcome that SHAP explains is determined entirely by the *prediction function* provided to the SHAP explainer. Let $f : \mathcal{X} \to \mathbb{R}^k$ denote the model output function defined over input features $\mathcal{X} \subseteq \mathbb{R}^d$, where k is the dimensionality of the output (e.g., number of classes). SHAP explanations are computed with respect to this function f, meaning that the computed Shapley values attribute contributions to the *output of* f.

$$\phi_j^{(f)}(x) \approx \text{contribution of feature } j \text{ to } f(x) \tag{9}$$

Thus, SHAP can be used not only to explain a model's prediction, but also any deterministic function derived from the model. This makes SHAP a flexible tool for explaining a variety of prediction targets, e.g., probabilities, logits, calibrated p-values, decision scores, confidence metrics, or structured outputs, by simply wrapping them in the function f. So, the *choice of function* f is what determines *what* is being explained, while SHAP decomposes this outcome into feature-level contributions under its additive game-theoretic framework.

To explain the p-values of conformal classifiers, we define a function $f_y(x) = p_y(x)$ for each class $y \in \mathcal{Y}$, where $p_y(x) \in [0, 1]$ is the conformal p-value. Thus, SHAP values are computed for each $f_y(x)$, explaining how features contribute to increasing or decreasing the plausibility of class y.

Below, we outline four other distinct strategies for applying SHAP explanations to outputs derived from conformal classification. Each formulation highlights a specific function of interest for which local feature attributions can be computed.

1. **Explaining class inclusion in the prediction set:**

   For each class $y \in \mathcal{Y}$, we define the binary function

   $$f_y(x) = \mathbb{1}\{p_y(x) > \epsilon\} \tag{10}$$

   where $p_y(x)$ is the conformal p-value for class y and ϵ is the significance threshold. SHAP values are computed for each $f_y(x)$, thereby explaining why class y was included (or not) in the prediction set for input x.

2. **Explaining the full prediction set as a structured output:**

   Let $\mathcal{S}(x) \subseteq \mathcal{Y}$ be the prediction set obtained by thresholding p-values. Each unique set $S \subseteq \mathcal{Y}$ is assigned an integer code $c_S \in \{0, \ldots, 2^{|\mathcal{Y}|} - 1\}$ using a binary encoding. Define the function

   $$f_{\text{set}}(x) = c_{\mathcal{S}(x)} \tag{11}$$

   and use SHAP to explain why $f_{\text{set}}(x)$ takes on a particular value. This enables the explanation of the entire prediction set as a categorical output.

3. **Explaining the confidence score:**

   Define the confidence function as:

$$f_{\text{conf}}(x) = 1 - \max_{y \in \mathcal{Y} \setminus \{y^*\}} p_y(x) \tag{12}$$

   where $y^* = \arg\max_{y \in \mathcal{Y}} p_y(x)$. This quantity measures how decisive the conformal prediction is. SHAP values explain which features contribute to a higher or lower confidence level in the prediction.

4. **Explaining the credibility score:**

   Define the credibility function as:

$$f_{\text{cred}}(x) = \max_{y \in \mathcal{Y}} p_y(x) \tag{13}$$

   which quantifies how well the input conforms to the most plausible class, or more generally how strange the instance is with respect to the calibration data. SHAP values reveal how individual features influence the credibility of the prediction.

3.3. Experimental setup

As the underlying model, we use a Random Forest classifier (Breiman, 2001) with 300 trees. We used Tree SHAP (Lundberg et al., 2020), with default settings as implemented in the SHAP Python package, to compute the feature attributions. To support the proof-of-concept nature of this study, we demonstrate our approach on three well-known benchmark datasets; *Iris*, *Wisconsin Breast Cancer*, and *Pima Diabetes*. These datasets have clearly defined features and targets, typically in the form of physical measurements, making both the problems and the input variables easy to interpret. For simplicity, we adopt a basic stratified hold-out scheme, allocating 60% of the data for training, 30% for calibration, and 10% for testing. All conformal classifiers are calibrated in a class-conditional manner.

4. Results.

We begin our analysis using the Iris dataset. This dataset contains four features (length and width of sepals and petals) of 50 samples each of three species of Iris (Setosa, Virginica, and Versicolor). As seen in Fig. 1, the Setosa class is linearly separable from the other two based on shorter and smaller petals.

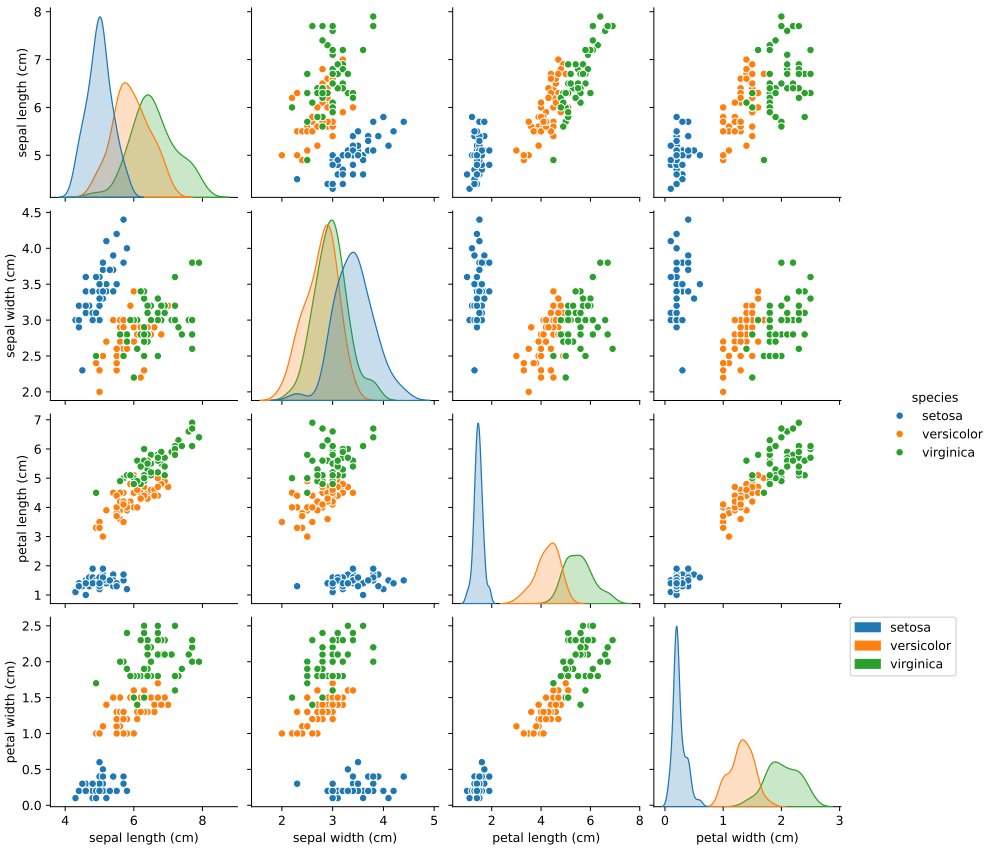

Figure 1: Iris. Pairwise feature plot.

In the first example, the random forest model assigns a predicted probability of 1.0 to the class Virginica. The model scores are then calibrated by the class-conditional conformal classifier, yielding the following p-values: $\{0.034, 0.035, 0.790\}$. Consequently, the conformal prediction is also Virginica (which is correct), and both the confidence (0.965) and the credibility (0.790) are high.

As shown in the SHAP waterfall plot in Fig. 2, the high values for petal width (in particular), petal length, and sepal length all support the prediction of Virginica. For this instance, the explanations are consistent across classes: the same feature values that increase the p-value for Virginica simultaneously decrease the p-values for Setosa and Versicolor.

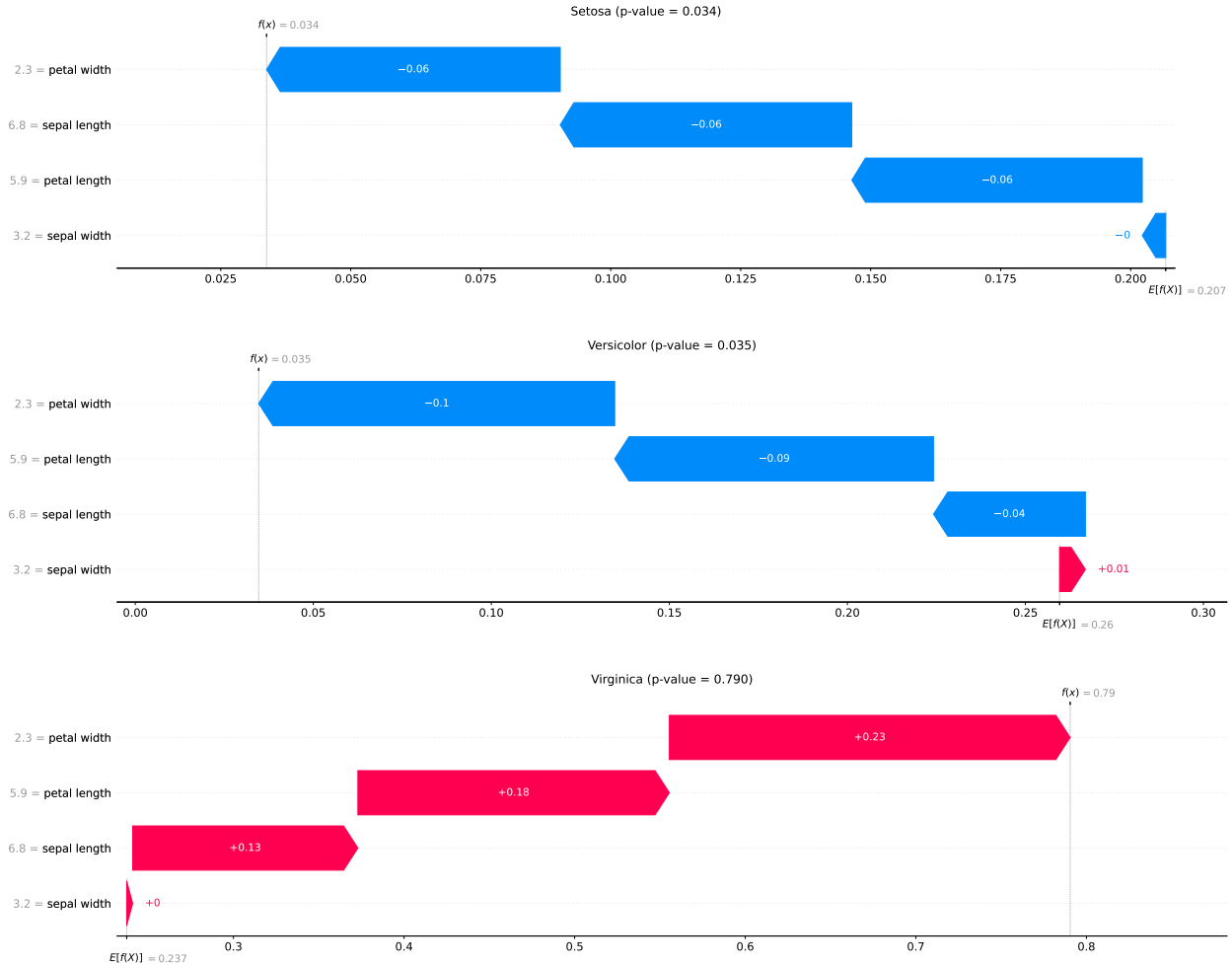

Figure 2: Iris instance 1. Explaining p-values

In the second example, the model outputs the class probabilities $\{0.000, 0.570, 0.430\}$, effectively ruling out Setosa but expressing uncertainty between Versicolor and Virginica. The conformal p-values, however, are $\{0.030, 0.096, 0.219\}$, resulting in a prediction shift from Versicolor to Virginica, which turns out to be the correct class. The confidence of the conformal predictor is relatively high (0.904), indicating strong support for rejecting Versicolor as well, but the low credibility (0.219) reveals that the instance is atypical compared to the calibration set.

Examining the SHAP waterfall plot in Fig. 3, we observe that all feature values contribute to lowering the p-value for Setosa. The large petal width increases the p-value for Virginica while decreasing it for Versicolor, supporting the selected prediction. Furthermore, the petal length decreases the p-values for both Virginica and Versicolor. This type of nuanced pattern can only be identified by using p-values instead of model scores, and explaining these separately for each class.

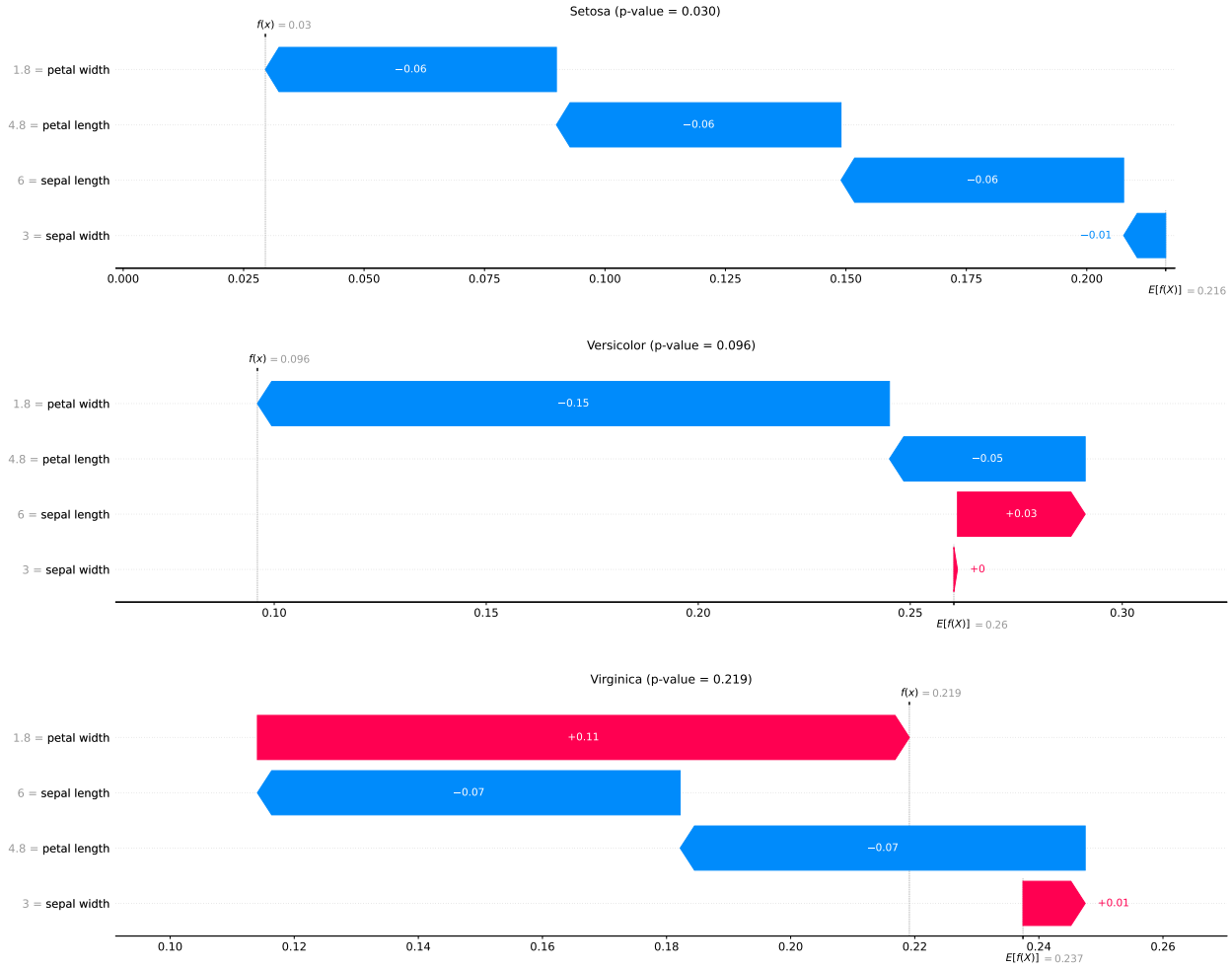

Figure 3: Iris instance 2. Explaining p-values

A SHAP beeswarm plot for class-wise conformal p-values provides a global explanation of how each input feature influences the p-value for a specific class across many instances. Features with higher absolute SHAP values have greater impact, and the direction (left or right) indicates whether a feature tends to decrease the p-value (pushing the class toward exclusion) or increase it (favoring inclusion). As shown in Fig. 4, the explanations align well with the known characteristics of the Iris classes: low feature values generally raise the p-values for Setosa but lower them for Versicolor and Virginica. Conversely, very high feature values, particularly for petal width, increase the p-value for Virginica while reducing it for Versicolor. Overall, the beeswarm plot offers a clear and intuitive summary of how feature values affect the p-values associated with each class.

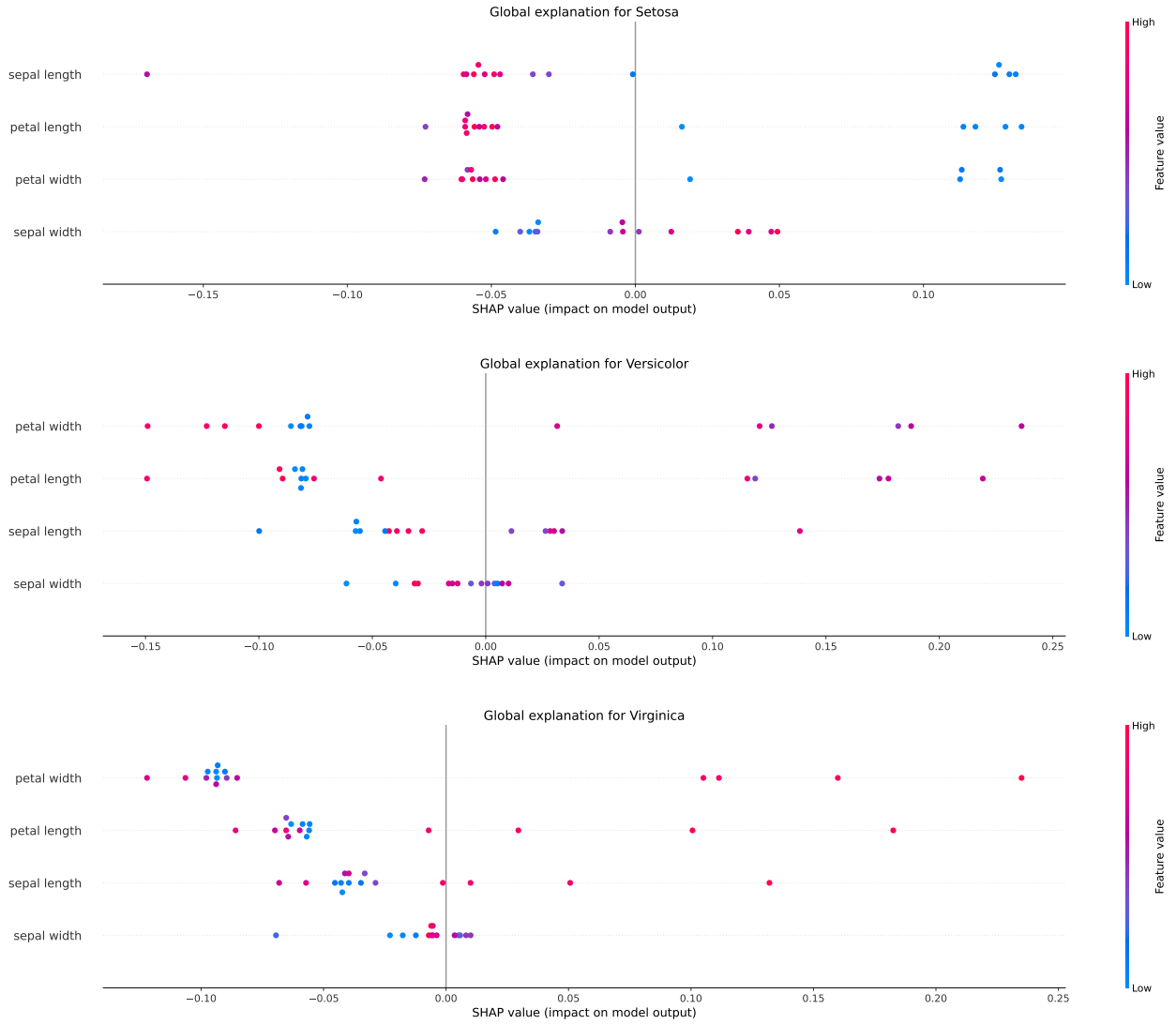

Figure 4: Iris global explanation.

Next, we present a few examples from the (original) Wisconsin Breast Cancer dataset. The task is to predict whether a given tumor is benign or malignant based on a set of features derived from the characteristics of individual cells. All feature values are integers ranging from 1 to 10, with higher values generally associated with an increased risk of malignancy.

In the first example, the model outputs scores $\{0.060, 0.940\}$, strongly (and correctly) favoring the Malignant class. The corresponding p-values are $\{0.015, 0.408\}$, making it possible to reject Benign at most common significance levels. While the confidence is very high (0.985), the credibility (0.408) is somewhat low, suggesting that this instance is not typical compared to the calibration set.

As shown in the waterfall plot in Fig. 5, all feature values contribute to lowering the p-value for Benign, clearly supporting its exclusion. For the Malignant class, however, the picture is more complex: some features values push the p-value up, while others, such as Bland Chromatin and Uniformity of Cell Size, push it down. Notably, the value for Uniformity of Cell Size lowers the p-values for both classes, illustrating the detailed insight that becomes possible when using p-values rather than raw model scores, and when inspecting the explanations separately for each class.

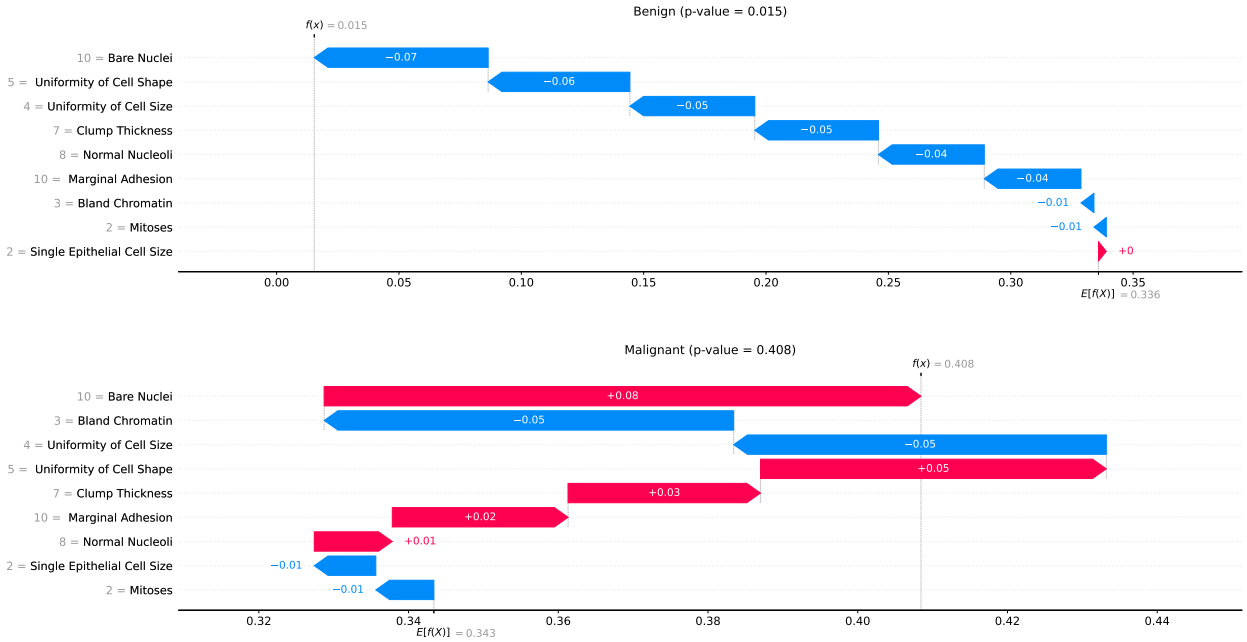

Figure 5: WBC instance. Explaining p-values

We now turn to explaining a different aspect of the prediction: the credibility. Specifically, the waterfall plot in Fig. 6 illustrates how the credibility of the instance is influenced by individual feature values. Consistent with the analysis above, we observe that Uniformity of Cell Size and Bland Chromatin, in particular, are key contributors to lowering the credibility. Since credibility reflects how similar an instance is to those in the calibration set, one interpretation is that these feature values are unusual, or at least atypical in combination with the others.

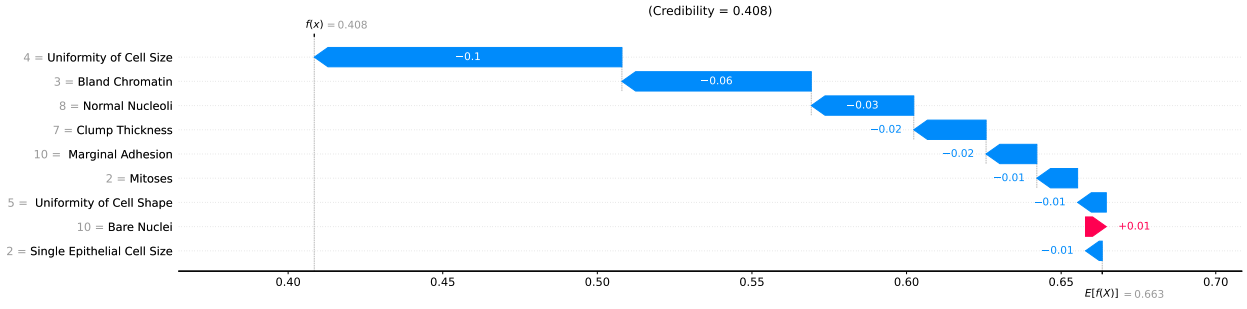

Figure 6: WBC instance. Explaining credibility

Next, we turn to explaining set inclusion, see Fig. 7. In this example, we use a significance level of $\epsilon = 0.05$, meaning that the Malignant class is included in the prediction set, while Benign is not. In this setup, the explained output is binary: 1 indicates inclusion, and 0 indicates rejection. While the SHAP explanations for set inclusion are naturally related to those for the underlying p-values, the application of the threshold introduces a shift in perspective.

For Benign, which is excluded, the explanation closely mirrors that of the class p-value: the same features that lowered the p-value also drive the rejection. However, for Malignant, which is

included, the SHAP values are now all positive. This may seem counterintuitive at first, as some of these features previously contributed to lowering the p-value. The key difference is that in the set inclusion explanation, the SHAP values describe how each feature contributes to crossing the threshold for inclusion, not how they affect the p-value itself. Even features that slightly lower the p-value may still support inclusion overall if the final value remains above the threshold. In this way, the inclusion explanation highlights which features made the class pass the decision boundary, offering a threshold-aware interpretation of the prediction.

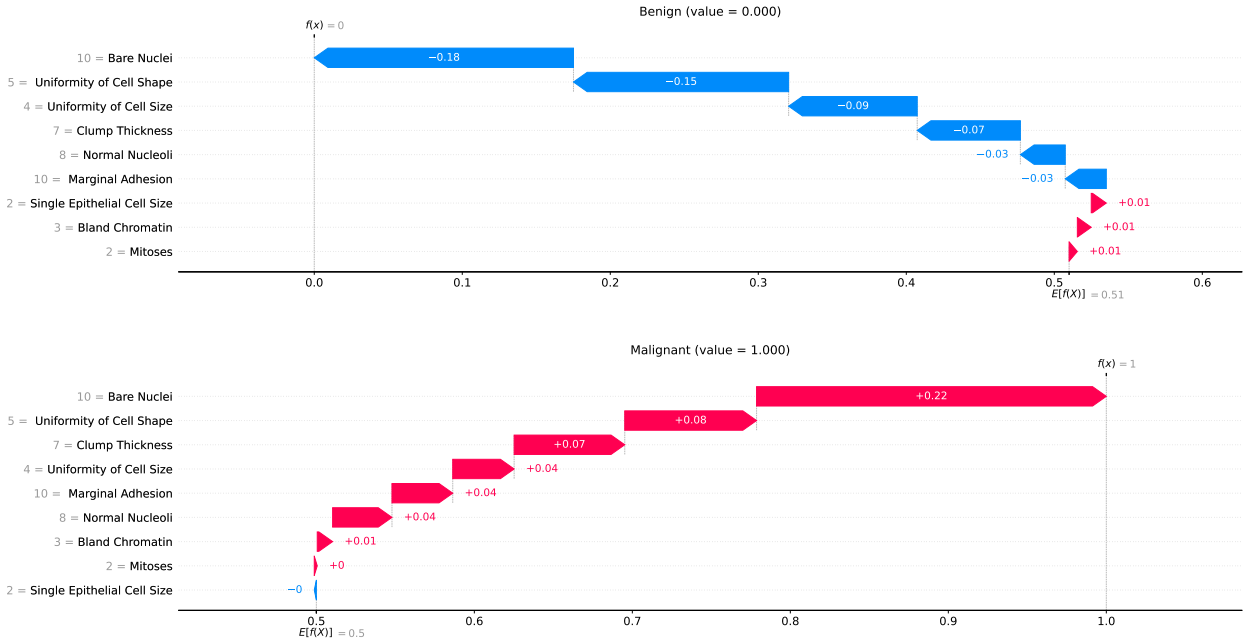

Figure 7: WBC instance. Explaining set inclusion. ϵ= 0.05

For the final set of examples, we use the Pima Indian Diabetes dataset. The goal of this dataset is to diagnostically predict whether a patient has diabetes, based on several clinical measurements. All patients are females aged 21 or older of Pima Indian heritage. As with the WBC dataset, higher feature values are generally associated with a higher risk, in this case of being diabetic.

In the instance under examination, the model outputs class probabilities of {0.537, 0.463}, leading to a classification of Non-Diabetic. However, class-conditional conformal calibration produces p-values of {0.180, 0.358}, resulting in a conformal prediction of Diabetic, which turns out to be correct. Nevertheless, the fact that neither class can be rejected at significance levels up to $\epsilon = 0.18$ clearly indicates that this is an uncertain prediction.

The SHAP explanation of the p-values in Fig. 8 reveals that, for the Non-Diabetic class, all feature values, except the low number of pregnancies, contribute to lowering the p-value from its expected baseline of 0.55, yet the resulting p-value remains relatively high (0.180). For the Diabetic class, which has a lower expected value (0.264), some feature values push the p-value up, while others pull it down.

In summary, these explanations make it clear that the conformal classifier is uncertain in this case, and identify which feature values contribute to that uncertainty.

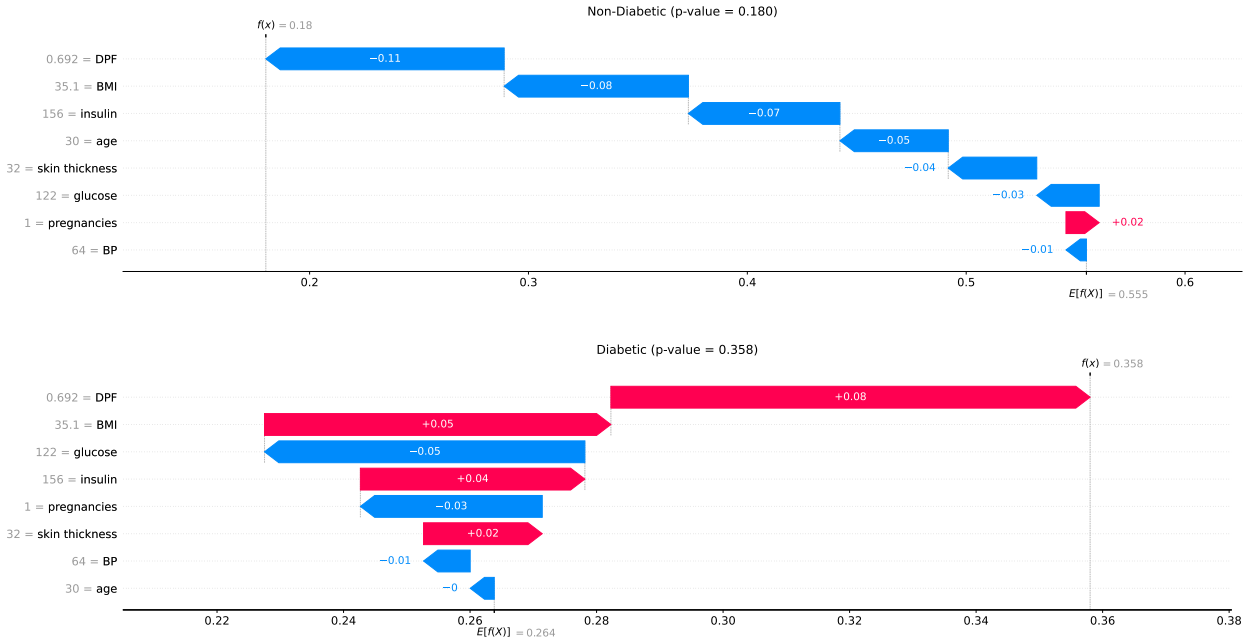

Figure 8: Diabetes instance. Explaining p-values

At first glance, the confidence might appear relatively high (0.820), while the credibility is somewhat low (0.358). However, a closer examination reveals that even the confidence value indicates a high degree of uncertainty compared to most other instances, see Fig. 9. As shown in the figure, all feature values contribute to reducing the confidence from a high baseline of 0.969 down to 0.820. In particular, the values for age and glucose, both highly relevant to the classification, are close to the population mean, offering limited discriminative power in this case.

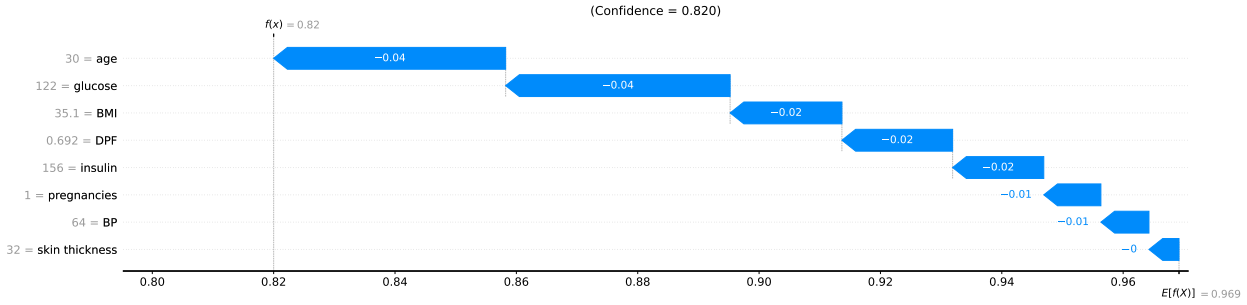

Figure 9: Diabetes instance. Explaining confidence

Finally, we directly explain the set prediction, which at $\epsilon = 0.05$ is {Non-Diabetic, Diabetic}. Similar to set inclusion, the explained output is 1.0, for the predicted set. As shown in Fig. 10, all feature values contribute to increasing the plausibility of this label set, with age and glucose emerging as the most influential features.

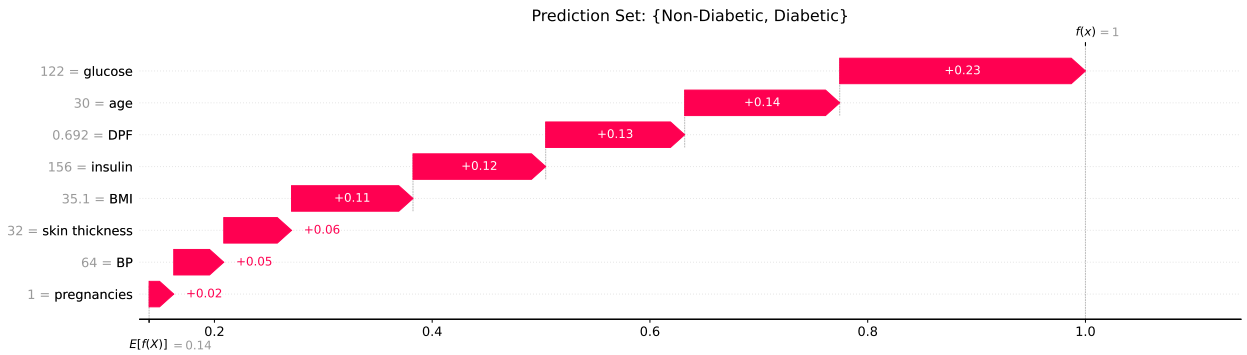

Figure 10: Diabetes instance. Explaining set prediction

5. Concluding remarks.

In this proof-of-concept study, we have proposed combining conformal classification with SHAP analysis to provide both local and global explanations of set predictions. The primary focus has been on explaining the p-values associated with each class, treating them as model outputs suitable for SHAP decomposition. This enables users to understand which classes were close to inclusion and how individual features contributed to the inclusion or exclusion of each label. Such class-specific, feature-level reasoning is difficult to achieve with standard softmax outputs, where explanations are constrained by the zero-sum nature of probability distributions.

By making the reasons for uncertainty explicit, conformal prediction combined with SHAP supports a more cautious, calibrated, and interpretable decision-making process. This is particularly valuable in high-stakes domains where errors must be explained and understood, not merely minimized.

While the main contribution of this paper is the SHAP-based explanation of p-values, we also outline how the same methodology can be extended to explain other conformal outputs, including set inclusion decisions, the full predicted label set, and derived measures such as confidence and credibility. Together, these targets form a versatile explanation framework for different aspects of conformal prediction.

To conclude, conformal prediction and SHAP together offer a powerful and extensible toolkit for interpretable uncertainty, enabling both fine-grained local explanations and global insight across complex predictive tasks.

Future work includes systematic empirical evaluations and extending the approach to conformal regression. In addition, the value of SHAP-enhanced conformal outputs should be assessed through user studies, either with real practitioners or simulated decision-makers.

Acknowledgements

The authors acknowledge the Swedish Knowledge Foundation, Jönköping University, and the industrial partners for financially supporting the research through the following projects: AFAIR (grant 20200223), PREMACOP (grant 20220187) and ETIAI (grant 20230036), as part of the research and education environment SPARK at Jönköping University, Sweden.

References

H. Abdelqader, E. Smirnov, M. Pont, and M. Geijselaers. Interpretable and reliable rule classification based on conformal prediction. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2022)*, pages 385–401. Springer, 2023.

Ernst Ahlberg, Susanne Winiwarter, Henrik Boström, Henrik Linusson, Tuve Löfström, Ulf Norinder Ulf Johansson, Ola Engkvist, Oscar Hammar, Claus Bendtsen, and Lars Carlsson. Using conformal prediction to prioritize compound synthesis in drug discovery. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 174–184. PMLR, 13–16 Jun 2017.

Amr Alkhatib, Henrik Boström, and Ulf Johansson. Assessing explanation quality by venn prediction. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, pages 42–54. PMLR, 2022. URL https://proceedings.mlr.press/v179/alkhatib22a.html.

Amr Alkhatib, Henrik Boström, Sofiane Ennadir, and Ulf Johansson. Approximating score-based explanation techniques using conformal regression. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 450–469. PMLR, 2023. URL https://proceedings.mlr.press/v204/alkhatib23a.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

David Cortés-Ciriano and Andreas Bender. Reliable prediction of bioactivity using conformal prediction and sparse data representations. *Journal of Cheminformatics*, 11(1):1–16, 2019.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. Application of conformal prediction in qsar. In *Artificial Intelligence Applications and Innovations*, pages 305–314. Springer, 2012.

Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. *arXiv preprint arXiv:2303.10694*, 2023.

Eyke Hüllermeier, Johannes Fürnkranz, and Eneldo Loza Mencía. Conformal rule-based multi-label classification. *arXiv preprint arXiv:2007.08145*, 2020.

Ulf Johansson, Rikard König, Tuve Löfström, and Henrik Boström. Evolved decision trees as conformal predictors. In *2013 IEEE Congress on Evolutionary Computation*, pages 1794–1801, 2013a. doi: 10.1109/CEC.2013.6557778.

Ulf Johansson, Henrik Linusson, Tuwe Löfström, and Henrik Boström. Conformal prediction using decision trees. In *Proceedings of the 13th IEEE International Conference on Data Mining*, pages 330–339. IEEE, 2013b.

Damjan Krstajic. Missed opportunities in large scale comparison of qsar and conformal prediction methods and their applications in drug discovery. *arXiv preprint arXiv:2001.07773*, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Tuwe Löfström, Helena Löfström, and Ulf Johansson. Calibrated explanations for multi-class. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 175–194. PMLR, 2024. URL https://proceedings.mlr.press/v230/lofstrom24a.html.

Tuwe Löfström, Helena Löfström, Ulf Johansson, Cecilia Sönströd, and Rudy Matela. Calibrated explanations for regression. *Machine Learning*, 114(4):100, February 2025. doi: 10.1007/s10994-024-06642-8. URL https://link.springer.com/10.1007/s10994-024-06642-8.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pages 417–431, September 2020 2020. doi: 10.1007/978-3-030-65965-3_28.

Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling for regulatory purposes. a transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling*, 54(6):1596–1603, 2014.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. *Tools in Artificial Intelligence*, 18:315–330, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

Jianfeng Sun, Lars Carlsson, Ernst Ahlberg, Ulf Norinder, Ola Engkvist, and Hongming Chen. Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *Journal of Chemical Information and Modeling*, 57(7):1591–1598, 2017.

Fredrik Svensson, Ulf Norinder, Henrik Boström, and Martin Eklund. Conformal regression for reliable prediction intervals. *Journal of Chemical Information and Modeling*, 59(3):1173–1182, 2019.

Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: A tree-based conformal prediction approach. *arXiv preprint arXiv:2404.19472*, 2024.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., 2005.