Proceedings of Machine Learning Research 329:1–24, 2026 Conformal and Probabilistic Prediction with Applications

# Performance Estimation in Hybrid Interpretable Models with Venn Predictors

**Alberto García-Galindo** AGARCIAGALI@UNAV.ES
**Marcos López-De-Castro** MLOPEZDECAS@UNAV.ES
**Ruben Armañanzas** RARMANANZAS@UNAV.ES
*Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Pamplona, Spain*
*TECNUN School of Engineering, Universidad de Navarra, Donostia-San Sebastián, Spain*
**Niclas Ståhl** NICLAS.STAHL@JU.SE
**Tuwe Löfström** TUWE.LOFSTROM@JU.SE
*Department of Computing, Jönköping University, Jönköping, Sweden*

**Editor:** Ernst Ahlberg, Ulf Johansson, Henrik Boström, Alberto Carlevaro, Johan Hallberg Szabadváry and Lars Carlsson

## Abstract

The design of hybrid interpretable models has recently emerged as a promising paradigm for the development of explainable artificial intelligence. This modeling framework relies on a collaborative scheme in which black-box and interpretable models are paired and cooperate to produce a final prediction. Intuitively, it assumes that there are some regions of the feature space where a black-box classifier can be replaced by an interpretable model without losing predictive performance. For hybrid interpretable models to be useful in practice, users should have statistical guarantees on their performance for a given transparency level (i.e., the fraction of samples delegated to the interpretable classifier). In this work, we propose the use of Venn prediction to construct reliable accuracy estimators for hybrid interpretable models in the absence of ground truth labels. In particular, we derive estimators for the marginal accuracy (i.e., for the hybrid model as a single predictor) and the component-conditional accuracy (i.e., for each of the internal components of the hybrid model). We demonstrate the benefits of our estimation methodology, consistently outperforming alternative approaches for six different datasets, with the conditional Venn predictor providing the most reliable component-conditional estimates.

**Keywords:** explainable artificial intelligence, hybrid interpretable models, Venn predictors, performance estimation

## 1. Introduction

Typical predictive modeling workflows often involve the comparison of different learning algorithms. Each algorithm, once fitted, leads to a single model instance that approximates the underlying relationship between a set of input variables and a target outcome. The performance of each model is then evaluated using proper performance metrics on unseen data to assess the generalization ability. In practice, we may be interested in different properties of the resulting model. The most obvious one is the predictive performance, but many others can be sought, especially when the predictive model is intended for use in high-risk applications (Eshete, 2021). In such settings, the development of explainable

models becomes a key requirement to understand why an algorithm produces a certain prediction and how the input features influence the outcome. This requirement of predictive models has gained attention from regulatory domains. For example, the EU's General Data Protection Regulation (GDPR) has called for the "right to explanation" (a right to inform individuals about the reasons behind an algorithm-informed decision) that requires human-understandable rationale of predictive models (European Parliament and Council of the European Union, 2016).

To address this challenge, Explainable Artificial Intelligence (XAI) has emerged as a field by itself dedicated to improving the comprehension and transparency of machine learning algorithms (Arrieta et al., 2020). This has been achieved mainly by explaining the underlying mechanisms of black-box systems, such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), or by designing interpretable-by-design models, such as rule-based classifiers (Wang et al., 2017). However, the former may not faithfully mimic the decision rationale of the original model, while the latter may sacrifice predictive performance. A potential solution to address these weaknesses is to consider a hybrid approach in which black-box and interpretable models are paired and cooperate to produce a final prediction (Ferry et al., 2024). Intuitively, a hybrid interpretable model delegates samples either to an interpretable white-box classifier, sacrificing predictive performance for explainability, or to a complex black-box classifier, with higher accuracy but losing the explanation. Under this scheme, the fraction of samples delegated to the white-box classifier defines the transparency of the hybrid predictive model. Consequently, as we increase the transparency of the hybrid model, its accuracy decreases, introducing a trade-off. For a hybrid model to be useful in practice, users should be able to anticipate its performance for a given transparency at inference time and support a safe deployment. Similarly, it should produce calibrated uncertainty estimates in order to employ them for decision-making. Unfortunately, off-the-shelf machine learning classifiers, including interpretable decision trees and complex neural networks, tend to generate poorly calibrated probabilities, which prevents hybrid models based on them from meeting these requirements (Provost and Domingos, 2003; Guo et al., 2017).

**Contribution**. In this work, we present procedures for inducing reliable hybrid predictive models whose accuracy-transparency trade-off can be estimated. Specifically, we propose the use of Venn prediction (Vovk et al., 2005) to calibrate a hybrid predictive model and induce a new predictor that yields well-calibrated probabilities. As a consequence of this calibration property, we show that the accuracy of the hybrid predictor can be precisely estimated at inference time, in the absence of ground truth labels. Specifically, we derive unbiased estimators for the marginal accuracy (i.e., without distinguishing whether the predictions were produced using either the white-box or the black-box model) and for the component-conditional accuracy (i.e., conditioning on the set of predictions assessed with each model). We theoretically demonstrate the validity of this procedure as long as the multiprobabilistic output of a Venn predictor is employed to produce an interval-valued estimate. Additionally, we empirically show, using six binary datasets, that when the multiprobabilistic output is merged into a single probability, the resulting estimators, while losing the guarantees, outperform those derived from uncalibrated and Platt-scaled classifiers.

**Outline**. The rest of this paper is organized as follows. Section 2 presents background on hybrid predictive models for XAI and briefly introduces Venn and Venn-Abers predictors. Section 3 formally describes the proposed procedures for estimating the performance of a hybrid interpretable model at inference time. In Section 4, we empirically evaluate the estimation quality of our contribution, comparing it with alternative approaches. Finally, Section 5 concludes the paper and discusses several directions of future work.

## 2. Background

### 2.1. Hybrid predictors for XAI

The design of hybrid predictive models (also referred to as companion models) has recently emerged as a promising framework in the context of XAI. Most of the proposed procedures are designed to increase the interpretability of machine learning classifiers without sacrificing predictive performance. Wang (2019) was the first to propose such a model, introducing a hybrid rule-set classifier that, given a pre-trained black-box classifier, induces a set of rules that act as an interpretable partial substitute on the subset of the input space where the black-box classifier is unnecessary. Only instances that are not covered by any rule are passed to the black-box classifier. One important drawback of this approach is that the user does not have explicit control over the transparency of the resulting hybrid model. This limitation was solved by Pan et al. (2020) by returning multiple hybrid models with different transparency levels using rule lists. Wang and Lin (2021) extended the scope of hybrid interpretable models to mixed schemes that included sparse linear models. Similarly, Li et al. (2026) proposed the integration of decision trees in the hybrid framework, achieving a better accuracy-transparency trade-off compared to previous works.

### 2.2. Venn and Venn-Abers predictors

Venn predictors (Vovk et al., 2005) are probabilistic predictors that return multiple probabilities for each label, with one of them being the valid one. These multiple probabilities can be converted into probability intervals, where the interval size conveys the confidence in the estimation. A computationally efficient version of Venn predictors is inductive Venn predictors, which can act as wrappers on top of any probabilistic predictor, referred to as the underlying model (Lambrou et al., 2015). To induce a Venn predictor in a binary setting, we consider a measurable feature space $\mathcal{X}$ and a binary response space $\mathcal{Y} = \{0, 1\}$. We assume the existence of a probabilistic (uncalibrated) classifier $f : \mathcal{X} \to [0, 1]$ that outputs the estimated likelihood of a binary outcome and an available set of $n$ independent and identically distributed (i.i.d.) calibration samples $\mathcal{D}_{\text{cal}} = \{(X_1, Y_1), ..., (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ drawn from an unknown data generation process $\mathcal{Q}$.

Central to the construction of a Venn predictor is the definition of a Venn taxonomy. Formally, a Venn taxonomy is a measurable function $K$ that defines the relation between a sample and a set of (other) samples which, in the inductive setting, is the training dataset $\mathcal{D}_{\text{train}}$. We define $K_i = K\big(\mathcal{D}_{\text{train}}, (X_i, Y_i)\big)$ as the category for the $i$-th calibration sample. Typically, the taxonomy is based on the underlying probabilistic classifier, where the estimated scores can be used to group samples into one of the categories. For a new test sample $(X_{n+1}, Y_{n+1})$ falling into a category, the predicted score is the relative frequency of

1s (i.e., positive labels) among all calibration samples in that category. By leveraging the i.i.d. assumption and by including the new sample in the computation, the predicted score is a calibrated probability. However, note that to compute the frequency of 1s, we would need to know $Y_{n+1}$ in advance, which is not possible at prediction time. For this reason, we tentatively consider each possible label $y \in \mathcal{Y} = \{0, 1\}$ for the unknown label $Y_{n+1}$ in the frequency computation, thereby producing two probability values. Formally, a Venn predictor outputs the multiprobabilistic prediction

$$P^y_{n+1} = \frac{y + \sum_{i=1}^{n} \mathbb{1}\{K_i = K^y_{n+1}\}\mathbb{1}\{Y_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{K_i = K^y_{n+1}\} + 1},$$

where $K^y_{n+1}$ is the category of the tentative test sample and $\mathbb{1}\{\cdot\}$ is the indicator function.

Regardless of the chosen taxonomy, the multiprobabilistic predictions $(P^0_{n+1}, P^1_{n+1})$ produced by a Venn predictor are perfectly calibrated by construction, where $P^0_{n+1} < P^1_{n+1}$. Intuitively, $P^0_{n+1}$ and $P^1_{n+1}$ are the predicted probabilities that $Y_{n+1} = 1$, and they can be seen as an interval-valued probabilistic prediction.

**Proposition 1** Let $\mathcal{D}_{\text{cal}} \cup (X_{n+1}, Y_{n+1})$ be a set of randomly drawn i.i.d. samples and $B \in \{0, 1\}$ a binary selector. Let $[P^0_{n+1}, P^1_{n+1}]$ be the multiprobabilistic prediction from a Venn predictor induced on $\mathcal{D}_{\text{cal}}$. There exists a selector $B$ such that

$$\mathbb{E}\big[Y_{n+1} \big| P^B_{n+1} = s\big] = s \quad \text{for all } s \in [0, 1] \text{ almost surely.}$$

**Proof** This is a well-established result of Venn prediction. For example, see Proposition 6.1 in Vovk et al. (2005). ■

In practice, Proposition 1 is usually interpreted as a guarantee on the prediction interval $[P^0_{n+1}, P^1_{n+1}]$. In particular, there exists a random variable $P_{n+1} \in [P^0_{n+1}, P^1_{n+1}]$ that is perfectly calibrated:

$$\mathbb{E}\big[Y_{n+1} \big| P_{n+1} = s\big] = s \quad \text{for all } s \in [0, 1] \text{ almost surely.}$$

A limitation of Venn predictors is that choosing an appropriate taxonomy can be challenging. One alternative is to use Venn-Abers predictors, in which the taxonomy is automatically induced through isotonic regression (Vovk and Petej, 2014). Isotonic regression is a piecewise-constant predictor $g : [0, 1] \to [0, 1]$ that partitions the score space into bins by minimizing the empirical mean squared error. For any sample assigned to a given bin, isotonic regression returns the relative frequency of positive labels within that bin. Thus, Venn-Abers predictors can be viewed as a special case of Venn predictors in which the taxonomy is determined by the isotonic regression fitted to the scores produced by the underlying model. In this way, they inherit the calibration guarantees of Venn predictors while remaining compatible with an inductive scheme.

Let $s_i = f(X_i)$ denote the score assigned by the underlying model to calibration sample $i$, and let $s_{n+1} = f(X_{n+1})$ denote the score assigned to the test sample. In practice, a Venn-Abers prediction can be computed by fitting two isotonic regressions, $g_0$ and $g_1$, on the augmented calibration sets

$$\{(s_1, y_1), \ldots, (s_n, y_n), (s_{n+1}, 0)\} \quad \text{and} \quad \{(s_1, y_1), \ldots, (s_n, y_n), (s_{n+1}, 1)\},$$

respectively. The resulting probability interval is then given by

$$[g_0(s_{n+1}), g_1(s_{n+1})].$$

## 3. Methods

### 3.1. Problem setup

We define a hybrid model as $f_{\text{hyb}} = \langle f_w, f_b \rangle$, where $f_w : \mathcal{X} \to [0,1]$ is an available white-box probabilistic classifier and $f_b : \mathcal{X} \to [0,1]$ is an available black-box probabilistic predictor. Previously, the design of hybrid interpretable models has been motivated by the assumption that there are some regions of the feature space where an interpretable model can be employed without trading off performance (Wang and Lin, 2021). With this in mind, a central component in the design of hybrid interpretable models is the definition of a delegation rule. In this work, we quantify the confidence of each sample using the white-box model, delegating those that fall below a predefined confidence threshold $\theta \in [0.5, 1]$ to the black-box classifier, formally

$$f_{\text{hyb}}(x) = \begin{cases} f_w(x), & \text{if } c_w(x) \geq \theta; \\ f_b(x), & \text{otherwise,} \end{cases} \tag{1}$$

where $c_w(x) = \max\{f_w(x), 1 - f_w(x)\}$ is the confidence in the prediction estimated with the white-box classifier. Under this scheme, the feature space is implicitly partitioned into two different component-wise subspaces, $\mathcal{X} = \mathcal{X}_w \cup \mathcal{X}_b$.

Our general objective is to estimate the accuracy of the hybrid interpretable model (1) by leveraging an available calibration dataset $\mathcal{D}_{\text{cal}}$. For a given hybrid predictor $f_{\text{hyb}}$, we compute the predicted label as $\hat{Y}_{n+1} = \mathbb{1}\{f_{\text{hyb}}(X_{n+1}) \geq 0.5\}$ and distinguish two different metrics: the *marginal accuracy* $A(f_{\text{hyb}})$ and the *component-conditional accuracy* $A_z(f_{\text{hyb}})$:

- **Marginal accuracy**. This metric evaluates the hybrid model as a single predictor, regardless of whether the predictions are produced by the white-box or the black-box component. Formally,
$$A(f_{\text{hyb}}) := \mathbb{E}\Big[\mathbb{1}\{Y_{n+1} = \hat{Y}_{n+1}\}\Big]. \tag{2}$$
- **Component-conditional accuracy**. This metric evaluates predictive performance after conditioning on the component that produced the prediction. In this way, it quantifies the accuracy separately within each delegated region of the feature space. Formally,
$$A_z(f_{\text{hyb}}) := \mathbb{E}\Big[\mathbb{1}\{Y_{n+1} = \hat{Y}_{n+1}\} \Big| X_{n+1} \in \mathcal{X}_z\Big] \text{ for all } z \in \{w, b\}. \tag{3}$$

These two notions of accuracy are useful in complementary scenarios. Marginal accuracy is the appropriate metric when the hybrid model is deployed as a single predictor and the primary objective is to characterize its overall predictive performance. By contrast, component-conditional accuracy is more informative when one wants to assess whether the interpretable component remains reliable on the samples delegated to it, or whether the black-box component preserves its expected performance on the more difficult cases.

### 3.2. Marginal accuracy estimation

We present a procedure for estimating the marginal accuracy of a hybrid interpretable model using Venn prediction. We begin by revisiting the output of Venn predictors. For a new sample $X_{n+1} \in \mathcal{X}$, let $[P^0_{n+1}, P^1_{n+1}]$ be the probability interval for $Y_{n+1} = 1$ and $\hat{Y}_{n+1} \in \{0, 1\}$ be the binary prediction. We introduce the confidence probability interval as follows:

$$[C^0_{n+1}, C^1_{n+1}] = \begin{cases} [P^0_{n+1}, P^1_{n+1}] & \text{if } \hat{Y}_{n+1} = 1; \\ [1 - P^1_{n+1}, 1 - P^0_{n+1}] & \text{if } \hat{Y}_{n+1} = 0. \end{cases}$$

Intuitively, this confidence interval represents our belief that the returned prediction is correct. For example, if we produce a probability interval $[P^0_{n+1}, P^1_{n+1}] = [0.20, 0.25]$ and return a negative prediction, then the interval $[1 - P^1_{n+1}, 1 - P^0_{n+1}] = [0.75, 0.80]$ is a confidence-calibrated output for producing a correct (negative) prediction.

**Lemma 1** Let $\mathcal{D}_{\text{cal}} \cup (X_{n+1}, Y_{n+1})$ be a set of randomly drawn i.i.d. samples, $B \in \{0, 1\}$ a binary selector, $[C^0_{n+1}, C^1_{n+1}]$ the confidence interval from a Venn predictor induced on $\mathcal{D}_{\text{cal}}$, and $\hat{Y}_{n+1}$ the binary prediction. There exists a selector $B$ such that $C^B_{n+1}$ is perfectly calibrated for the correct classification $T_{n+1} = \mathbb{1}\{Y_{n+1} = \hat{Y}_{n+1}\}$:

$$\mathbb{E}\big[T_{n+1} \big| C^B_{n+1} = s\big] = s, \forall s \in [0, 1].$$

**Proof** This proof follows from the law of iterated expectations and Proposition 1. ■

From a confidence-calibrated classifier, Kivimäki et al. (2025) demonstrated that the average confidence computed on a set of $m$ i.i.d. samples $\mathcal{D}_{\text{test}} = \{(X_1, Y_1), ..., (X_m, Y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ drawn from the same data-generating process $\mathcal{Q}$, formally

$$\bar{C}_{\text{test}} = \frac{1}{m} \sum_{i=1}^{m} C_i,$$

is an unbiased estimator of the model accuracy.

We leverage this result to construct an estimator for the marginal accuracy of the hybrid predictor. To this end, we employ the white-box classifier $f_w$ and the black-box classifier $f_b$ to set a hybrid model $f_{\text{hyb}}$ as in (1) for a given confidence threshold $\theta \in [0.5, 1]$. Then, we employ $f_{\text{hyb}}$ to compute the estimated scores on $\mathcal{D}_{\text{cal}}$ and induce a Venn-Abers predictor using the standard procedure. Here, we consider the following taxonomy

$$K\big(\mathcal{D}_{\text{train}}, (x_i, y_i)\big) = g(f_{\text{hyb}}(x_i)), \tag{4}$$

where the category is given by the isotonic regression $g$ fitted on the scores estimated by $f_{\text{hyb}}$. That is, two samples are assigned to the same category only if they share the same score bin.

**Proposition 2** Let $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ be a set of randomly drawn i.i.d. samples, $B \in \{0, 1\}$ a binary selector, and $C^B_i$ be the confidence prediction for the $i$-th test sample from a Venn predictor induced on $\mathcal{D}_{\text{cal}}$ with the taxonomy (4). There exists a selector $B$ such that $\bar{C}^B_{\text{test}} = \frac{1}{m} \sum_{i=1}^{m} C^B_i$ is an unbiased estimator of the marginal accuracy (2).

**Proof** Consider a random test sample $(X_{n+1}, Y_{n+1})$ and the corresponding score computed using the hybrid model, $S_{n+1} = f_{\text{hyb}}(X_{n+1})$. Next, consider the calibration dataset $\mathcal{D}_{\text{cal}}$ and the corresponding scores computed using the hybrid model, $\{S_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{f_{\text{hyb}}(X_i)\}_{i\in\mathcal{D}_{\text{cal}}}$. Note that $S_{n+1}$ and the elements of $\{S_i\}_{i\in\mathcal{D}_{\text{cal}}}$ are i.i.d., so the calibration guarantees of Venn prediction hold. The rest of the proof follows by Lemma 1 and Theorem 2 in Kivimäki et al. (2025). ■

### 3.3. Component-conditional accuracy estimation

In some cases, we may be interested in providing a reliable accuracy estimate for each component of the hybrid interpretable model. However, a reliable marginal accuracy estimate does not rule out a scenario where we might be reporting an overconfident estimate for the white-box model while underestimating the performance of the black-box classifier, or vice versa.

Importantly, the validity of the marginal accuracy estimator in Proposition 2 is a consequence of the calibration property of Venn prediction. Therefore, to design a reliable estimator for the component-conditional accuracy, we need to ensure the estimated scores are calibrated conditionally on the internal component of the hybrid interpretable model. To this end, we propose the following Venn taxonomy

$$K\big(\mathcal{D}_{\text{train}}, (x_i, y_i)\big) = \begin{cases} g_w(f_w(x_i)), & \text{if } c_w(x_i) \geq \theta; \\ g_b(f_b(x_i)), & \text{otherwise,} \end{cases} \tag{5}$$

where $g_w$ and $g_b$ are component-wise isotonic regressions, each fitted only on the calibration samples that are delegated to the white-box classifier and the black-box classifier, respectively. That is, two samples are assigned to the same category only if they share both the same score bin and the same internal component of the hybrid predictor.

This simple modification is enough to preserve the calibration guarantees of Venn prediction within each component. In other words, the estimated confidence scores are not only calibrated marginally, but also conditioned on whether the prediction was delegated to the white-box classifier or to the black-box classifier.

**Proposition 3** Let $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ be a set of randomly drawn i.i.d. samples, $B \in \{0, 1\}$ a binary selector, and $C_i^B$ be the confidence prediction for the $i$-th test sample from a Venn predictor induced on $\mathcal{D}_{\text{cal}}$ with the taxonomy (5). Consider the set of test samples delegated to each component $z \in \{w, b\}$, $\mathcal{D}_{\text{test},z} = \{i \in [m] : X_i \in \mathcal{X}_z\}$ of size $m_z = |\mathcal{D}_{\text{test},z}|$. There exists a selector $B$ such that $\bar{C}^B_{\text{test},z} = \frac{1}{m_z}\sum_{i\in\mathcal{D}_{\text{test},z}} C_i^B$ is an unbiased estimator of the component-conditional accuracy (3).

**Proof** Consider a random test sample $(X_{n+1}, Y_{n+1})$ delegated to component $z \in \{w, b\}$ and the corresponding score $S_{n+1} = f_z(X_{n+1})$. Next, consider the calibration samples delegated to component $z$, $\mathcal{D}_{\text{cal},z} = \{i \in [n] : X_i \in \mathcal{X}_z\}$, with the corresponding scores $\{S_i\}_{i\in\mathcal{D}_{\text{cal},z}}$. The calibration and test samples delegated to component $z$ are i.i.d., so the calibration guarantees of Venn prediction hold within this component. The rest of the proof follows from Lemma 1 and Theorem 2 in Kivimäki et al. (2025). ■

### 3.4. Estimators derived from multiprobabilistic predictions

The results above build upon the theoretical calibration guarantees of Venn prediction. That is, they hold for (at least) one of the reported probability distributions. We now show that an interval-valued estimator for the marginal accuracy can be constructed by averaging the lower and upper confidence outputs of the Venn predictor. Formally, this estimator is defined as

$$\left[\bar{C}^0_{\text{test}}, \bar{C}^1_{\text{test}}\right], \tag{6}$$

where

$$\bar{C}^0_{\text{test}} = \frac{1}{m}\sum_{i=1}^{m} C^0_i, \qquad \bar{C}^1_{\text{test}} = \frac{1}{m}\sum_{i=1}^{m} C^1_i.$$

**Corollary 1** Under the setting of Proposition 2, the interval-valued estimator (6) from the Venn outputs satisfies

$$\mathbb{E}[\bar{C}^0_{\text{test}}] \leq A(f_{\text{hyb}}) \leq \mathbb{E}[\bar{C}^1_{\text{test}}]. \tag{7}$$

**Proof** By the definition of the Venn output, $C^0_i \leq C^B_i \leq C^1_i$ for every $i$-th test sample. Then, by linearity of summation, $\bar{C}^0_{\text{test}} \leq \bar{C}^B_{\text{test}} \leq \bar{C}^1_{\text{test}}$. The rest of the proof follows by taking expectations and Proposition 2. ■

An analogous interval can be constructed for the component-conditional accuracy by considering the taxonomy (5). In this case, the interval-valued estimator is defined as

$$\left[\bar{C}^0_{\text{test},z}, \bar{C}^1_{\text{test},z}\right], \tag{8}$$

where

$$\bar{C}^0_{\text{test},z} = \frac{1}{m_z}\sum_{i\in\mathcal{D}_{\text{test},z}} C^0_i, \qquad \bar{C}^1_{\text{test},z} = \frac{1}{m_z}\sum_{i\in\mathcal{D}_{\text{test},z}} C^1_i.$$

**Corollary 2** Under the setting of Proposition 3, the interval-valued estimator (8) from the Venn outputs satisfies

$$\mathbb{E}[\bar{C}^0_{\text{test},z}] \leq A_z(f_{\text{hyb}}) \leq \mathbb{E}[\bar{C}^1_{\text{test},z}].$$

**Proof** The proof follows the same argument as that of Corollary 1 using Proposition 3. ■

These intervals bound the estimator associated with the valid Venn selectors for every test set, and their expected limits cover the corresponding accuracy. Importantly, they provide valuable information for decision-making. For instance, the average of the lower confidence outputs can be interpreted as a risk-averse estimate of the accuracy that the resulting hybrid predictor is guaranteed, on average, to achieve at inference time. Therefore, it serves as a proxy for the level of transparency that can be ensured in practice.

In practice, it may be useful to produce a single confidence prediction for each sample and derive a point estimate for the hybrid predictor's accuracy. This can be done by aggregating the lower and the upper Venn outputs into a confidence scalar using the central point of the prediction interval,

$$C^*_{n+1} = \frac{C^0_{n+1} + C^1_{n+1}}{2}.$$

Consequently, we can compute an empirical point estimator for the marginal accuracy by replacing $\bar{C}^B_{\text{test}}$ in Proposition 2 with the average of the central points of the Venn prediction intervals,

$$\bar{C}^*_{\text{test}} = \frac{1}{m}\sum_{i=1}^{m}\frac{C_i^0 + C_i^1}{2}. \tag{9}$$

Similarly, a point estimate for the component-conditional accuracy can be derived by replacing $\bar{C}^B_{\text{test},z}$ in Proposition 3 with

$$\bar{C}^*_{\text{test},z} = \frac{1}{m_z}\sum_{i\in\mathcal{D}_{\text{test},z}}\frac{C_i^0 + C_i^1}{2}. \tag{10}$$

These point estimators lose the theoretical guarantees of the interval-valued estimators above. However, the results of our experiments, presented in the next section, empirically show that they can be used to provide reliable performance estimates.

## 4. Experiments

We conducted an experimental evaluation to demonstrate how the proposed method can be employed to provide reliable accuracy estimates for a hybrid interpretable model. To this end, we define two independent experiments. Experiment 1 focused on the marginal accuracy estimation task, whereas Experiment 2 was devoted to providing reliable component-conditional accuracy estimates.

### 4.1. Experimental setting

To construct the hybrid interpretable models, we employed shallow decision trees as white-box classifiers and XGBoost predictors as black-box classifiers. For XGBoost, we fixed the default hyperparameter configuration. For the shallow decision trees, we fixed a maximum depth to four in order to avoid over-complexity and ensure human comprehension. To train the algorithms, we used scikit-learn (Pedregosa et al., 2011) and XGBoost Python libraries, respectively. For this hybrid model configuration, we compared four different settings:

- Uncalibrated models (**Uncal**). Here, we employ off-the-shelf white-box and black-box classifiers to build a hybrid model and compute the average confidence on the test set.
- Platt scaling method (**Platt**). The hybrid model (1) is calibrated with Platt scaling using the CalibratedClassifierCV method from scikit-learn (Pedregosa et al., 2011). Then, the resulting confidence scores on the test set are employed in an identical way as for the uncalibrated models.
- Marginal Venn predictor (**Marg. Venn**). This time, the hybrid model (1) is calibrated using Venn prediction considering the taxonomy (4). Recall that this setting is identical to inducing a standard Venn-Abers predictor with the calibration scores produced by the hybrid model. To this end, we employed the venn-abers library (Petej, 2024). The point estimates are computed by taking the average confidence on the test set, whereas the interval-valued estimates are computed by taking the average lower and upper values of the prediction interval.

- Conditional Venn Predictor (**Cond. Venn**). In this case, the hybrid model (1) is calibrated using Venn prediction considering the taxonomy (5). That is, two different Venn-Abers predictors are induced: one with the calibration samples delegated to the white-box model and one with the calibration samples delegated to the black-box model. As before, we employed the venn-abers library (Petej, 2024). The point and interval-valued accuracy estimates are computed in an identical way as for the marginal Venn predictor.

In our experiments, the testing protocol was a repeated stratified hold-out validation with 100 independent runs of a 75/25% random split into training and test sets, respectively. For calibration purposes, each training dataset was further partitioned into 67% proper training samples and 33% calibration samples. To explore different points of the accuracy-transparency trade-off, we sorted samples, both in the calibration set (when it was used) and the test set, according to the confidence estimated with the white-box classifier. Then, the subset of the most confident samples was delegated to the white-box model, considering a range of transparency levels from 10% to 90%.

To evaluate each setting, we computed the average estimated accuracy and the average observed accuracy on test sets across runs. Then, we evaluated the differences between both metrics, yielding an estimation error. In Experiment 1, all test samples were considered, whereas in Experiment 2, we independently computed the estimation error on each of the internal components of the hybrid model.

In total, we employed six different binary classification datasets: adult, eeg, magic, phoneme, phishing, and ailerons, available at the UCI (Kelly et al., 2025) and the OpenML (Vanschoren et al., 2014) repositories. Table 1 presents a summary of the datasets.

Table 1: Description of datasets, including number of samples ($n$), number of features ($p$), fraction of positive samples, and source.

| Dataset | $n$ | $p$ | Positive frac. | Source |
|---|---|---|---|---|
| adult | 48,842 | 14 | 0.239 | UCI |
| eeg | 14,980 | 14 | 0.449 | OpenML |
| magic | 19,020 | 11 | 0.352 | UCI |
| phishing | 11,055 | 30 | 0.557 | UCI |
| phoneme | 5,404 | 5 | 0.293 | OpenML |
| ailerons | 13,750 | 40 | 0.424 | OpenML |

### 4.2. Baseline evaluation

To establish a baseline for the construction of hybrid interpretable models, we initially present the accuracy of both the shallow decision trees and the XGBoost classifiers in Table 2 for each setting. As expected, the XGBoost classifiers outperformed the decision trees in all settings, with accuracy improvements ranging from 2.7% (on adult) to 23% (on eeg). On average, the uncalibrated classifiers yielded slightly better accuracy than the Platt-scaled models and the Venn predictors, especially for the XGBoost.

Table 2: Baseline accuracy.

| | Decision tree | | | XGBoost | | |
|---|---|---|---|---|---|---|
| Dataset | Uncal | Platt | Venn | Uncal | Platt | Venn |
| adult | .845 | .845 | .845 | .872 | .870 | .870 |
| eeg | .695 | .691 | .691 | .929 | .921 | .920 |
| magic | .817 | .818 | .818 | .880 | .878 | .877 |
| phishing | .918 | .918 | .918 | .970 | .966 | .966 |
| phoneme | .792 | .791 | .791 | .894 | .885 | .884 |
| ailerons | .840 | .839 | .839 | .883 | .881 | .881 |
| **Mean** | **.818** | **.817** | **.817** | **.905** | **.900** | **.900** |

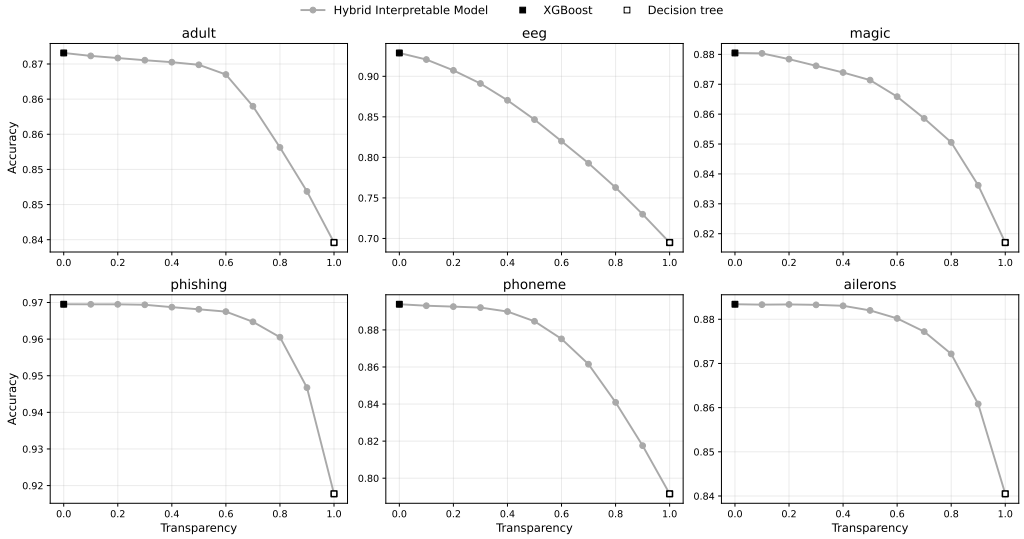


Figure 1: Accuracy-transparency trade-off. The black squares represent the black-box classifiers (XGBoost), the white squares represent the white-box classifiers (decision tree), and the gray circles represent several hybrid interpretable models for varying transparencies.

We now show the average accuracy-transparency trade-off for the six datasets in Figure 1. Because the curves were similar for all settings, we only present the trade-off using the uncalibrated hybrid models. The horizontal axis represents the transparency, and the vertical axis represents the accuracy. Note that, as more samples were delegated to the white-box classifiers, the accuracy of the resulting hybrid models decreased. This was the common pattern in all cases, which aligns with the logic of hybrid interpretable models. Interestingly, the trade-off curves were different depending on the dataset. For adult, phoneme, phishing, and ailerons, about 50% of the samples could be delegated to the decision tree without losing the performance of the XGBoost. In these datasets, our delegation rule was able to identify easy-to-predict regions of the feature space, demonstrating the benefits of the hybrid modeling framework. However, this was not the case for magic and eeg, where the performance quickly degraded even for low transparency levels and the decision tree was not able to detect samples that were easy to classify.

### 4.3. Experiment 1: Marginal accuracy estimation

Next, we present the marginal accuracy estimation results for each of the considered settings. Figures 2 and 3 depict both the realized and the estimated performance for each setting. For the Venn approaches, the solid curves represent the central point estimates, whereas the shaded regions represent the interval-valued accuracy estimates obtained by separately averaging the lower and upper Venn confidence outputs.

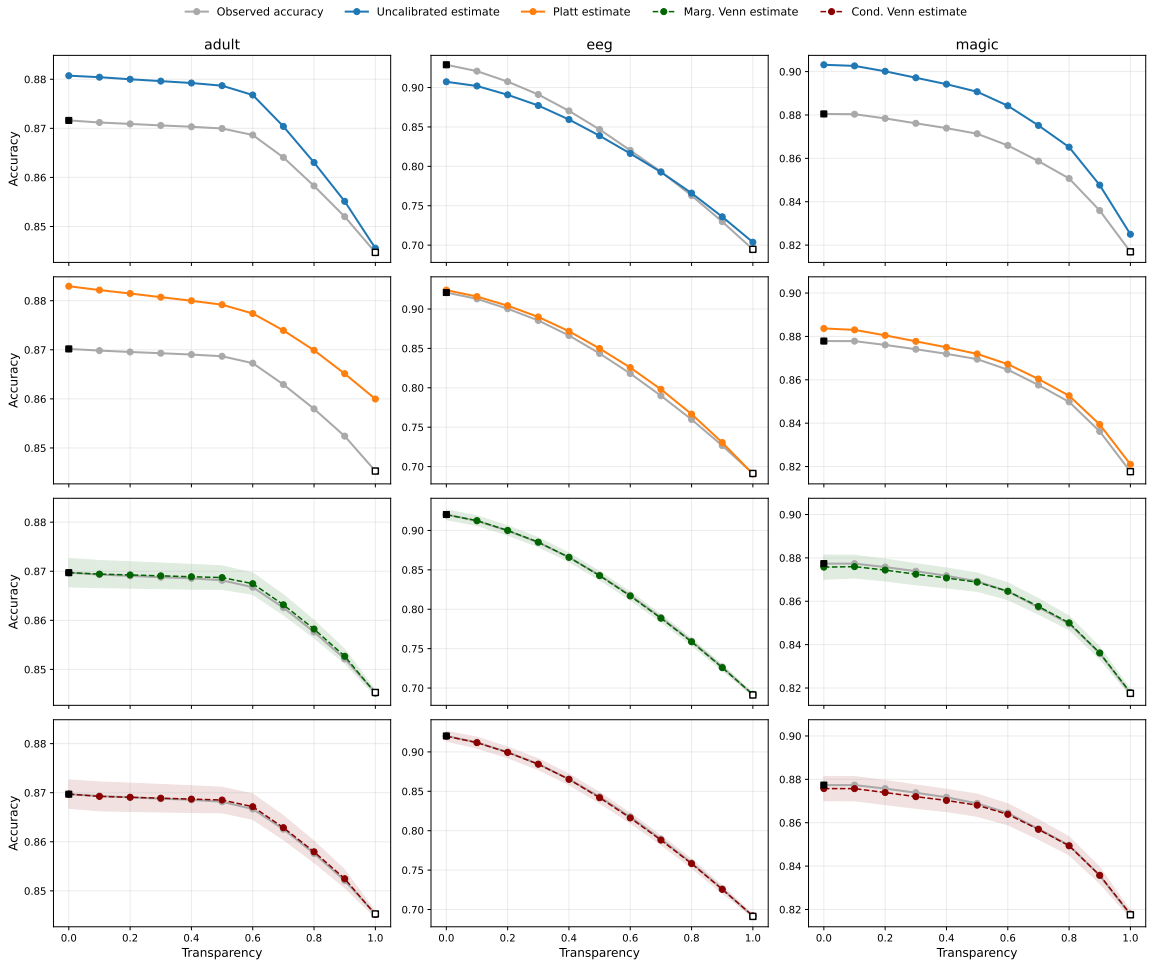


Figure 2: Marginal accuracy estimation results on adult, eeg, and magic. The shaded regions represent the interval-valued estimates obtained by averaging the lower and upper Venn confidence outputs.

Focusing on the estimation quality of the uncalibrated confidences, we observe that the produced accuracy estimates overestimated the observed accuracy on average. Interestingly, the quality of the estimates increased for higher transparency levels. That is, as the hybrid models were more similar to the decision trees, they produced more reliable probabilities. This was the case, for example, of the ailerons dataset. Here, the estimation error delegating all the samples to the XGBoost was about 3.5%, gradually decreasing to less than 1%

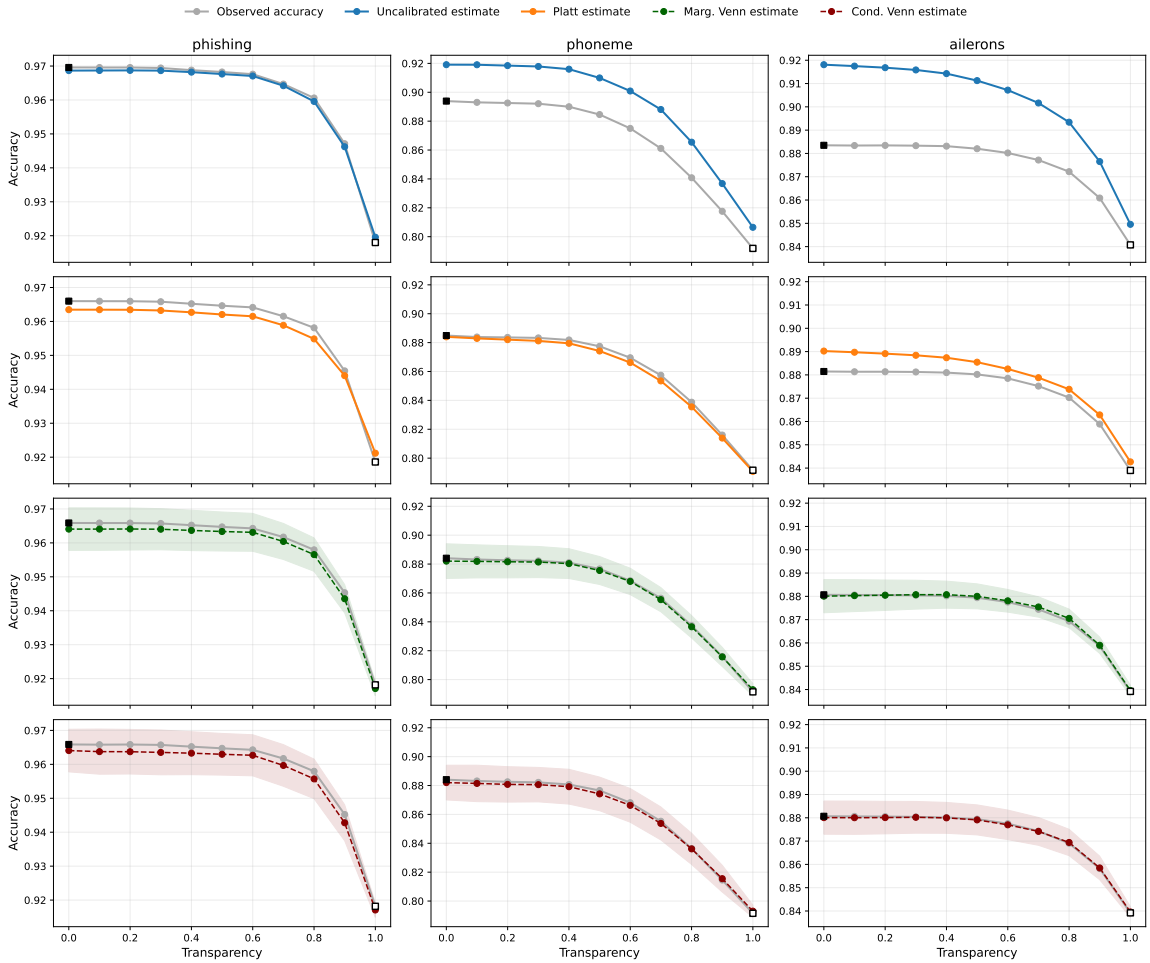


Figure 3: Marginal accuracy estimation results on phishing, phoneme, and ailerons. Other details are as in Figure 2.

when employing the decision tree. The estimates produced using Platt scaling partially solved this problem, but they still yielded unreliable performance estimates in some cases. On the adult dataset, this approach deteriorated the original estimation quality of the uncalibrated models, yielding an overestimation error of 1.5%. Conversely, the accuracy estimates reported with the proposed Venn approaches were consistently more reliable, emerging as the best-performing alternatives. Both Venn approaches yielded similar point estimates for the marginal accuracy, regardless of the chosen taxonomy. In line with the theoretical guarantees, the average accuracy was consistently covered by the corresponding Venn interval across the complete transparency range. Interestingly, the size of the intervals decreased as more samples were delegated to the decision tree.

Table 3 reports the signed and absolute estimation errors averaged across datasets for each transparency level. Consistent with the figures above, uncalibrated models and Platt scaling produced unreliable estimates, with average errors of 1.5% and 0.5%, respectively. These methods systematically overestimated the performance across all transparency levels.

In contrast, marginal and conditional Venn predictors generated accuracy estimates that deviated from the observed accuracies by only 0.1% across all transparency levels. This indicates a clear advantage over competitors, with estimates that are effectively near-perfect. Additional point estimation results for some particular transparency levels can be found in Tables A1-A4. The corresponding interval-valued Venn estimates are presented in Tables A5-A8, showing that the average accuracy lay within the corresponding interval in every case.

Table 3: Signed and absolute marginal estimation errors averaged across datasets, by transparency level.

| | Uncal | | Platt | | Marg. Venn | | Cond. Venn | |
|---|---|---|---|---|---|---|---|---|
| Transp. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. |
| 10% | .012 | .019 | .004 | .005 | -.001 | .001 | -.001 | .001 |
| 20% | .012 | .018 | .004 | .005 | -.001 | .001 | -.001 | .001 |
| 30% | .012 | .017 | .004 | .005 | .000 | .001 | -.001 | .001 |
| 40% | .012 | .016 | .003 | .005 | .000 | .001 | -.001 | .001 |
| 50% | .012 | .015 | .003 | .005 | .000 | .001 | -.001 | .001 |
| 60% | .012 | .014 | .003 | .005 | .000 | .001 | -.001 | .001 |
| 70% | .012 | .013 | .003 | .005 | .000 | .001 | -.001 | .001 |
| 80% | .011 | .012 | .003 | .005 | .000 | .001 | -.001 | .001 |
| 90% | .009 | .009 | .003 | .005 | .000 | .001 | .000 | .001 |
| **Mean** | **.012** | **.015** | **.003** | **.005** | **.000** | **.001** | **-.001** | **.001** |

### 4.4. Experiment 2: Component-conditional accuracy estimation

We now turn to the component-conditional accuracy estimation results. In this case, we independently evaluated the performance estimates of the white-box component and the black-box component of the hybrid interpretable model. Figures 4 and 5 illustrate both the realized and the estimated performance using each approach. Again, uncalibrated and Platt-scaled confidences failed to provide a reliable component-conditional accuracy estimation. In the case of the uncalibrated models, most of the estimation errors are observed on the samples assessed with the XGBoost. For the Platt scaling method, we see a similar pattern across datasets: as the transparency increased, the estimation error on the samples assessed with the XGBoost increased as well. Conversely, the estimation quality on the samples delegated to the decision tree improved for the highest transparency level. With this approach, we observe that providing a reliable marginal accuracy estimate does not protect us against uneven estimation quality on any of the components of the hybrid model. Focusing on the eeg dataset, we found that the Platt-scaled confidences provided near-perfect marginal estimates but performed substantially worse when restricted to individual internal classifiers of the hybrid model, with estimation errors of up to 5%.

The differences between the observed and the estimated component-conditional accuracy using the marginal Venn predictors were substantially reduced for all datasets. However, the produced intervals failed to contain the conditional accuracies in many settings. This behavior is expected because the marginal Venn predictor is calibrated for the hybrid model

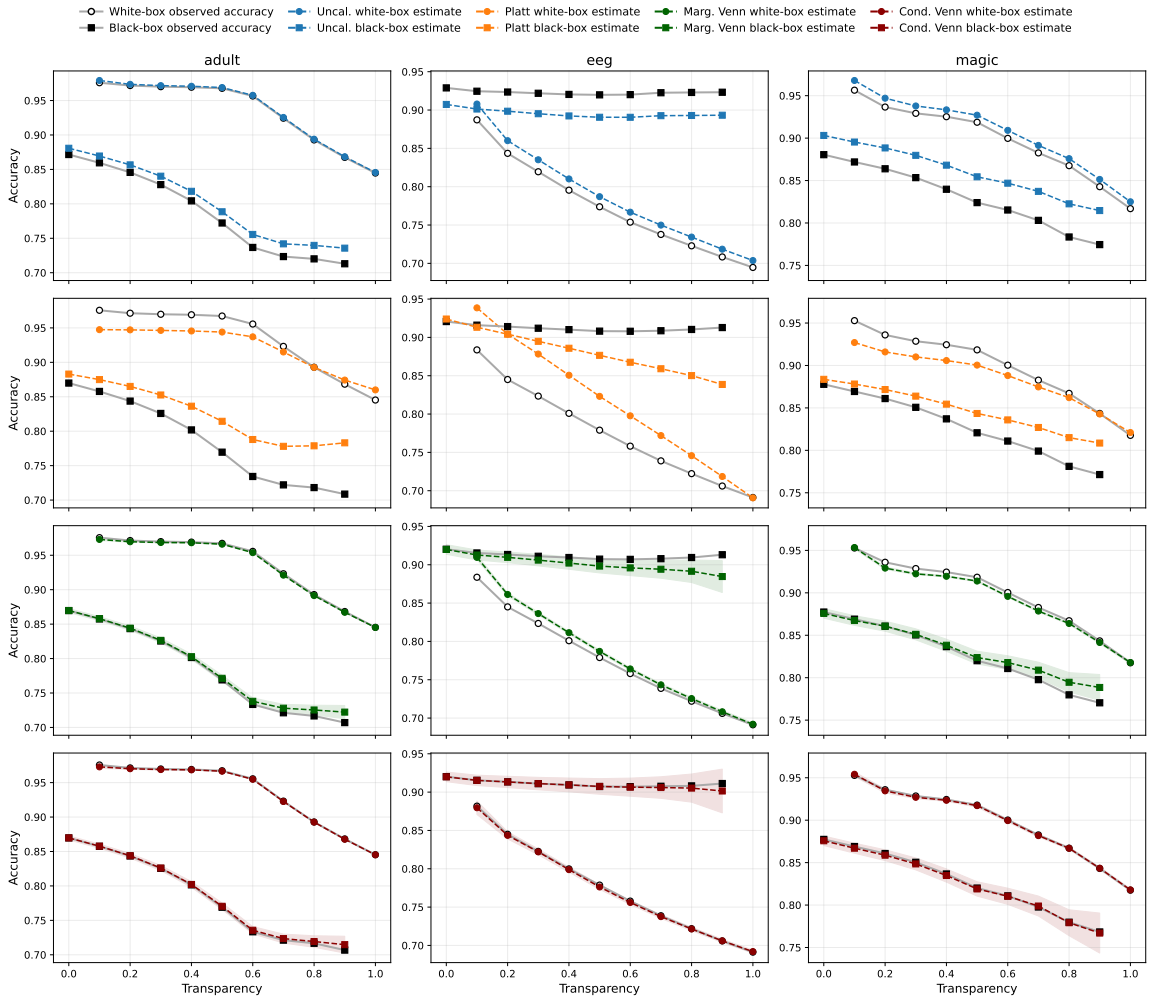


Figure 4: Component-conditional accuracy estimation results on adult, eeg, and magic. Other details are as in Figure 2.

as a whole and therefore provides no component-conditional calibration guarantees. By contrast, the interval-valued estimates of the conditional Venn predictors consistently covered the accuracy of each component in all cases. This approach provided precise accuracy estimates for all the considered transparency levels and outperformed competitors on every dataset.

Table 4 presents the component-conditional estimation errors averaged across all datasets. Focusing on the white-box component evaluation, we see that uncalibrated models and the marginal Venn approach performed relatively well for all the considered transparency levels. On average, these approaches yielded 0.7% and 0.5% estimation errors, respectively. Conversely, Platt scaling failed to provide reliable accuracy estimates for the samples assessed with the interpretable part of the hybrid model, producing a mean absolute error of 2.3%. The conditional Venn approach emerged as the best approach, clearly outperforming the competitors regardless of the transparency and delivering precise accuracy estimates.

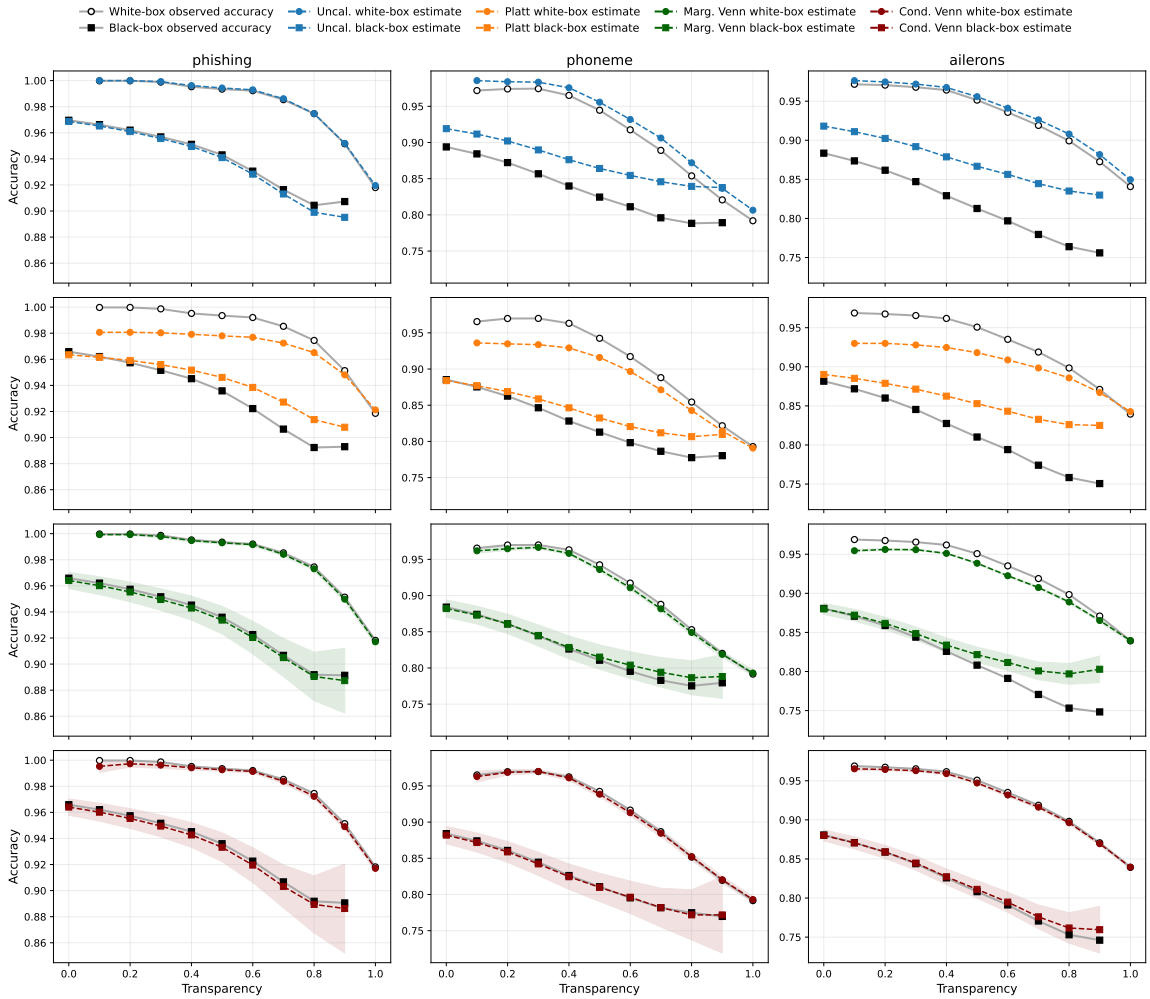


Figure 5: Component-conditional accuracy estimation results on phishing, phoneme, and ailerons. Other details are as in Figure 2.

Regarding the results on the black-box component of the hybrid model, uncalibrated and Platt-scaled classifiers produced very unreliable estimates, yielding a 2.9% estimation error on average. Marginal Venn also failed to estimate the accuracy of the samples delegated to the black-box component, especially for the higher transparency levels. For 70%, 80% and 90% transparency levels, this approach delivered 1.3%, 1.6%, and 2.2% estimation errors, respectively. Finally, the accuracy estimates produced with the conditional Venn predictors emerged again as the best ones. This method had a mean absolute estimation error of only 0.2% on average, without showing any systematic tendency to overestimate or underestimate the observed accuracy. It should be noted, however, that the conditional Venn estimates were slightly less accurate at the 90% transparency level. This deterioration can be attributed to the small number of calibration samples delegated to the black-box component, which also resulted in wider interval-valued estimates than at the other transparency levels. Additional component-conditional results for some particular transparency levels can

be found in Tables A9-A12, whereas the corresponding interval-valued Venn estimates are shown in Tables A13-A16. From these results, we observe that the Venn intervals for the white-box component were substantially narrower than those for the black-box component. The marginal Venn intervals often failed to cover the corresponding component-conditional accuracies. By contrast, the conditional interval-valued estimates empirically covered the true accuracy of each component in all cases.

Table 4: Signed and absolute component-conditional estimation errors averaged across datasets, by transparency level. Left block: white-box component; right block: black-box component.

| | White-box component | | | | | | | | Black-box component | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uncal | | Platt | | Marg. Venn | | Cond. Venn | | Uncal | | Platt | | Marg. Venn | | Cond. Venn | |
| Transp. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. | Sig. | Abs. |
| 10% | .009 | .009 | -.014 | .033 | .001 | .008 | -.002 | .002 | .012 | .020 | .006 | .007 | -.001 | .001 | -.001 | .001 |
| 20% | .007 | .007 | -.013 | .033 | -.002 | .007 | -.002 | .002 | .013 | .022 | .008 | .011 | .000 | .002 | -.001 | .001 |
| 30% | .007 | .007 | -.013 | .032 | -.001 | .006 | -.001 | .001 | .015 | .024 | .011 | .017 | .000 | .002 | -.001 | .001 |
| 40% | .006 | .006 | -.013 | .030 | -.002 | .005 | -.001 | .001 | .016 | .026 | .015 | .023 | .001 | .004 | -.001 | .001 |
| 50% | .007 | .007 | -.012 | .027 | -.003 | .005 | -.002 | .002 | .018 | .029 | .018 | .029 | .002 | .006 | .000 | .002 |
| 60% | .007 | .007 | -.009 | .022 | -.003 | .005 | -.002 | .002 | .020 | .031 | .021 | .034 | .005 | .009 | .001 | .002 |
| 70% | .008 | .008 | -.006 | .017 | -.003 | .005 | -.001 | .001 | .022 | .033 | .023 | .040 | .007 | .013 | .001 | .002 |
| 80% | .008 | .008 | -.003 | .010 | -.003 | .004 | -.001 | .001 | .024 | .036 | .025 | .045 | .010 | .016 | .000 | .003 |
| 90% | .008 | .008 | .001 | .006 | -.002 | .002 | -.001 | .001 | .024 | .038 | .026 | .051 | .011 | .022 | .002 | .007 |
| **Mean** | **.007** | **.007** | **-.009** | **.023** | **-.002** | **.005** | **-.001** | **.001** | **.018** | **.029** | **.017** | **.029** | **.004** | **.008** | **.000** | **.002** |

## 5. Conclusions

Hybrid interpretable models combine interpretable and black-box predictors to balance transparency and predictive performance. In this work, we have introduced a methodology for estimating the accuracy-transparency trade-off for hybrid interpretable models at inference time. The proposed methodology derives reliable accuracy estimators from the output of Venn predictors. More specifically, we introduced estimators for two complementary metrics: the marginal accuracy of the hybrid model, viewed as a single predictor, and the component-conditional accuracy, which evaluates the white-box and black-box components separately on the samples delegated to each of them.

Through an experimental study on six binary classification datasets, we have shown that the proposed estimators outperform estimators based on uncalibrated and Platt-scaled probabilities. Notably, the performance estimates using the conditional Venn predictors, where the internal component is integrated into the taxonomy definition, were the most precise for the component-conditional accuracy.

Future work should include extensive experiments using more datasets and different machine learning algorithms for assembling the hybrid models. Additionally, it would be interesting to explore how this approach could be extended to performance metrics other than accuracy and to other learning tasks, including regression, classification with a reject option, and fair learning.

## Acknowledgments

A. García-Galindo, M. López-De-Castro, and R. Armañanzas were supported by the Gobierno de Navarra through the ANDIA 2021 program (grant no. 0011-3947-2021-000023) and the ERA PerMed JTC2022 PORTRAIT project (grant no. 0011-2750-2022-000000). N. Ståhl and T. Löfström were supported by AFAIR (grant no. 20200223) and ETIAI (grant no. 20230036), as part of the research and education environment SPARK at Jönköping University, Sweden.

## References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

Birhanu Eshete. Making machine learning trustworthy. *Science*, 373(6556):743–744, 2021. doi: 10.1126/science.abi5052.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), 2016.

Julien Ferry, Gabriel Laberge, and Ulrich Aïvodji. Learning hybrid interpretable models: Theory, taxonomy, and methods. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=XzaSGIStXP.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository. https://archive.ics.uci.edu, 2025.

Juhani Kivimäki, Jukka K Nurminen, Jakub Białek, and Wojtek Kuberski. Confidence-based estimators for predictive performance in model monitoring. *Journal of Artificial Intelligence Research*, 82:209–240, 2025. doi: 10.1613/jair.1.16709.

Antonis Lambrou, Ilia Nouretdinov, and Harris Papadopoulos. Inductive Venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74(1):181–201, 2015. doi: 10.1007/s10472-014-9420-z.

Yifan Li, Shuhan Qi, Lei Cui, Chao Xing, Lei Zhang, and Xuan Wang. Interpret when possible: A tree-based hybrid framework for interpretable classification. *Big Data Mining and Analytics*, 9(1):263–283, 2026. doi: 10.26599/BDMA.2025.9020055.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Danqing Pan, Tong Wang, and Satoshi Hara. Interpretable companions for black-box models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2444–2454. PMLR, 26–28 Aug 2020.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ivan Petej. venn-abers: Venn-Abers calibration for probabilistic classifiers, 2024. URL https://github.com/ip200/venn-abers.

Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine learning*, 52(3):199–215, 2003. doi: 10.1023/A:1024099825458.

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*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Vladimir Vovk and Ivan Petej. Venn-Abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 829–838. AUAI Press, 2014. ISBN 9780974903910.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Cham, 1 edition, 2005. doi: 10.1007/b106715.

Tong Wang. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*, pages 6505–6514. PMLR, 2019.

Tong Wang and Qihang Lin. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *Journal of Machine Learning Research*, 22(137): 1–38, 2021.

Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37, 2017. URL http://jmlr.org/papers/v18/16-003.html.

## Appendix A. Additional results

### A.1. Marginal accuracy: point estimates

| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .871 | .880 | .869 | .881 | .869 | .869 | .869 | .869 |
| eeg | .891 | .877 | .885 | .890 | .885 | .885 | .885 | .884 |
| magic | .876 | .897 | .874 | .878 | .874 | .872 | .874 | .872 |
| phishing | .969 | .969 | .966 | .963 | .966 | .964 | .965 | .964 |
| phoneme | .892 | .918 | .883 | .881 | .882 | .881 | .882 | .881 |
| ailerons | .883 | .916 | .881 | .888 | .880 | .881 | .880 | .880 |

Table A1: Marginal accuracy estimation results: 30% transparency.

| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .870 | .879 | .869 | .879 | .868 | .869 | .868 | .869 |
| eeg | .847 | .839 | .844 | .850 | .843 | .843 | .843 | .842 |
| magic | .871 | .891 | .870 | .872 | .869 | .869 | .869 | .868 |
| phishing | .968 | .968 | .965 | .962 | .965 | .963 | .965 | .963 |
| phoneme | .885 | .910 | .877 | .874 | .877 | .876 | .877 | .874 |
| ailerons | .882 | .911 | .880 | .885 | .879 | .880 | .879 | .879 |

Table A2: Marginal accuracy estimation results for 50% transparency.

| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .864 | .870 | .863 | .874 | .863 | .863 | .863 | .863 |
| eeg | .793 | .793 | .790 | .798 | .790 | .789 | .789 | .788 |
| magic | .859 | .875 | .858 | .860 | .857 | .858 | .857 | .857 |
| phishing | .965 | .964 | .962 | .959 | .962 | .960 | .962 | .960 |
| phoneme | .861 | .888 | .857 | .853 | .856 | .855 | .855 | .854 |
| ailerons | .877 | .902 | .875 | .879 | .874 | .875 | .874 | .874 |

Table A3: Marginal accuracy estimation results for 70% transparency.

| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .852 | .855 | .852 | .865 | .852 | .853 | .852 | .852 |
| eeg | .730 | .736 | .727 | .731 | .727 | .726 | .726 | .726 |
| magic | .836 | .848 | .836 | .839 | .836 | .836 | .836 | .836 |
| phishing | .947 | .946 | .945 | .944 | .945 | .944 | .945 | .943 |
| phoneme | .818 | .837 | .816 | .814 | .816 | .816 | .815 | .816 |
| ailerons | .861 | .877 | .859 | .863 | .859 | .859 | .858 | .858 |

Table A4: Marginal accuracy estimation results for 90% transparency.

### A.2. Marginal accuracy: interval-valued Venn estimates

| | Marg. Venn | | | | Cond. Venn | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .869 | .866 | .872 | .006 | .869 | .866 | .872 | .006 |
| eeg | .885 | .879 | .891 | .012 | .885 | .878 | .891 | .013 |
| magic | .874 | .868 | .877 | .009 | .874 | .867 | .877 | .010 |
| phishing | .966 | .958 | .970 | .012 | .966 | .957 | .970 | .013 |
| phoneme | .882 | .871 | .892 | .021 | .882 | .869 | .893 | .024 |
| ailerons | .880 | .875 | .887 | .012 | .880 | .873 | .887 | .014 |

Table A5: Interval-valued marginal accuracy estimates for 30% transparency.

| | Marg. Venn | | | | Cond. Venn | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .868 | .866 | .871 | .005 | .868 | .866 | .871 | .005 |
| eeg | .843 | .837 | .848 | .011 | .843 | .835 | .848 | .013 |
| magic | .869 | .865 | .873 | .008 | .869 | .863 | .873 | .010 |
| phishing | .965 | .958 | .969 | .011 | .965 | .957 | .969 | .012 |
| phoneme | .877 | .866 | .885 | .019 | .877 | .863 | .886 | .023 |
| ailerons | .879 | .875 | .885 | .010 | .879 | .873 | .886 | .013 |

Table A6: Interval-valued marginal accuracy estimates for 50% transparency.

| | Marg. Venn | | | | Cond. Venn | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .863 | .861 | .865 | .004 | .863 | .860 | .865 | .005 |
| eeg | .789 | .784 | .793 | .009 | .789 | .782 | .794 | .012 |
| magic | .857 | .854 | .861 | .007 | .857 | .852 | .861 | .009 |
| phishing | .962 | .955 | .966 | .011 | .962 | .954 | .966 | .012 |
| phoneme | .856 | .847 | .864 | .017 | .855 | .842 | .865 | .023 |
| ailerons | .874 | .871 | .880 | .009 | .874 | .868 | .880 | .012 |

Table A7: Interval-valued marginal accuracy estimates for 70% transparency.

| | Marg. Venn | | | | Cond. Venn | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .852 | .851 | .854 | .003 | .852 | .851 | .854 | .003 |
| eeg | .727 | .722 | .730 | .008 | .726 | .721 | .730 | .009 |
| magic | .836 | .833 | .839 | .006 | .836 | .832 | .839 | .007 |
| phishing | .945 | .939 | .948 | .009 | .945 | .937 | .948 | .011 |
| phoneme | .816 | .809 | .822 | .013 | .815 | .806 | .825 | .019 |
| ailerons | .859 | .855 | .863 | .007 | .858 | .853 | .864 | .011 |

Table A8: Interval-valued marginal accuracy estimates for 90% transparency.

### A.3. Component-conditional accuracy: point estimates

| | White-box | | | | | | | | Black-box | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .970 | .972 | .970 | .946 | .970 | .969 | .970 | .969 | .828 | .840 | .826 | .853 | .825 | .826 | .825 | .826 |
| eeg | .819 | .835 | .823 | .878 | .823 | .836 | .823 | .822 | .922 | .895 | .912 | .895 | .911 | .906 | .911 | .911 |
| magic | .929 | .938 | .929 | .910 | .929 | .922 | .929 | .927 | .853 | .880 | .851 | .864 | .850 | .851 | .850 | .848 |
| phishing | .999 | .999 | .999 | .980 | .999 | .998 | .998 | .996 | .957 | .956 | .952 | .956 | .952 | .950 | .952 | .949 |
| phoneme | .974 | .983 | .970 | .934 | .970 | .967 | .970 | .970 | .857 | .890 | .846 | .859 | .845 | .845 | .845 | .842 |
| ailerons | .968 | .972 | .966 | .928 | .966 | .956 | .965 | .963 | .847 | .892 | .845 | .871 | .844 | .849 | .844 | .845 |

Table A9: Component-conditional accuracy estimation results for 30% transparency.

| | White-box | | | | | | | | Black-box | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .968 | .969 | .967 | .944 | .967 | .966 | .967 | .967 | .772 | .788 | .770 | .814 | .769 | .771 | .769 | .770 |
| eeg | .774 | .787 | .779 | .823 | .779 | .787 | .779 | .776 | .920 | .891 | .908 | .877 | .907 | .898 | .907 | .907 |
| magic | .919 | .927 | .918 | .900 | .918 | .914 | .918 | .917 | .824 | .854 | .821 | .843 | .820 | .824 | .820 | .819 |
| phishing | .993 | .994 | .994 | .978 | .994 | .993 | .994 | .993 | .943 | .941 | .936 | .946 | .936 | .934 | .936 | .933 |
| phoneme | .945 | .956 | .943 | .916 | .943 | .936 | .942 | .939 | .825 | .864 | .813 | .832 | .811 | .815 | .811 | .810 |
| ailerons | .951 | .956 | .951 | .918 | .951 | .938 | .951 | .947 | .813 | .867 | .810 | .853 | .808 | .822 | .808 | .811 |

Table A10: Component-conditional accuracy estimation results for 50% transparency.

| | White-box | | | | | | | | Black-box | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .924 | .925 | .923 | .915 | .923 | .921 | .923 | .923 | .723 | .742 | .722 | .778 | .721 | .728 | .721 | .724 |
| eeg | .738 | .750 | .739 | .772 | .739 | .743 | .739 | .738 | .923 | .893 | .909 | .859 | .908 | .894 | .908 | .906 |
| magic | .883 | .891 | .883 | .875 | .883 | .879 | .883 | .882 | .803 | .837 | .799 | .827 | .798 | .809 | .798 | .799 |
| phishing | .985 | .986 | .985 | .972 | .985 | .984 | .985 | .984 | .916 | .913 | .906 | .927 | .907 | .905 | .907 | .903 |
| phoneme | .889 | .906 | .888 | .871 | .888 | .882 | .887 | .885 | .796 | .846 | .786 | .812 | .783 | .794 | .782 | .782 |
| ailerons | .919 | .926 | .919 | .898 | .919 | .907 | .919 | .916 | .780 | .844 | .774 | .833 | .771 | .801 | .771 | .776 |

Table A11: Component-conditional accuracy estimation results for 70% transparency.

| | White-box | | | | | | | | Black-box | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | | Uncal. | | Platt | | Marg. Venn | | Cond. Venn | |
| Dataset | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. | Acc. | Est. |
| adult | .867 | .868 | .868 | .874 | .868 | .867 | .868 | .868 | .713 | .736 | .709 | .783 | .707 | .722 | .707 | .715 |
| eeg | .708 | .719 | .706 | .719 | .706 | .708 | .706 | .706 | .923 | .893 | .913 | .839 | .913 | .885 | .911 | .901 |
| magic | .843 | .851 | .843 | .843 | .843 | .841 | .843 | .843 | .775 | .815 | .771 | .809 | .770 | .789 | .768 | .767 |
| phishing | .952 | .952 | .951 | .948 | .951 | .950 | .951 | .949 | .907 | .895 | .893 | .908 | .892 | .887 | .891 | .886 |
| phoneme | .821 | .837 | .822 | .814 | .820 | .819 | .820 | .820 | .789 | .838 | .780 | .810 | .779 | .788 | .769 | .772 |
| ailerons | .873 | .882 | .871 | .867 | .871 | .865 | .871 | .869 | .756 | .830 | .751 | .825 | .748 | .803 | .746 | .759 |

Table A12: Component-conditional accuracy estimation results for 90% transparency.

### A.4. Component-conditional accuracy: interval-valued Venn estimates

| | Marg. Venn | | | | | | | | Cond. Venn | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White-box | | | | Black-box | | | | White-box | | | | Black-box | | | |
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | <u>.970</u> | .968 | .969 | .001 | .825 | .823 | .830 | .007 | .970 | .968 | .970 | .002 | .825 | .822 | .830 | .008 |
| eeg | <u>.823</u> | .835 | .838 | .003 | .911 | .899 | .914 | .015 | .823 | .818 | .826 | .008 | .911 | .903 | .919 | .016 |
| magic | <u>.929</u> | .921 | .923 | .002 | .850 | .845 | .858 | .013 | .929 | .925 | .929 | .004 | .850 | .841 | .856 | .015 |
| phishing | .999 | .997 | .999 | .002 | .952 | .941 | .958 | .017 | .999 | .994 | .999 | .005 | .952 | .941 | .958 | .017 |
| phoneme | <u>.970</u> | .965 | .968 | .003 | .845 | .830 | .860 | .030 | .970 | .967 | .972 | .005 | .845 | .826 | .858 | .032 |
| ailerons | <u>.966</u> | .955 | .957 | .002 | .844 | .840 | .857 | .017 | .964 | .962 | .964 | .002 | .844 | .836 | .854 | .018 |

Table A13: Interval-valued component-conditional accuracy estimates for 30% transparency. Accuracies not covered by the corresponding Venn interval are underlined.

| | Marg. Venn | | | | | | | | Cond. Venn | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White-box | | | | Black-box | | | | White-box | | | | Black-box | | | |
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .967 | .966 | .966 | .000 | .769 | .767 | .776 | .009 | .967 | .966 | .967 | .001 | .769 | .766 | .775 | .009 |
| eeg | .779 | .785 | .789 | .003 | .907 | .889 | .907 | .018 | .779 | .773 | .779 | .006 | .907 | .897 | .918 | .021 |
| magic | .918 | .913 | .915 | .003 | .820 | .816 | .831 | .015 | .918 | .915 | .919 | .004 | .820 | .811 | .827 | .016 |
| phishing | .994 | .992 | .994 | .002 | .936 | .923 | .944 | .021 | .994 | .991 | .994 | .002 | .936 | .922 | .944 | .022 |
| phoneme | .943 | .934 | .939 | .005 | .811 | .798 | .832 | .034 | .942 | .935 | .942 | .007 | .811 | .790 | .829 | .039 |
| ailerons | .951 | .937 | .940 | .003 | .808 | .812 | .831 | .019 | .949 | .946 | .949 | .003 | .808 | .800 | .822 | .022 |

Table A14: Interval-valued component-conditional accuracy estimates for 50% transparency. Accuracies not covered by the corresponding Venn interval are underlined.

| | Marg. Venn | | | | | | | | Cond. Venn | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White-box | | | | Black-box | | | | White-box | | | | Black-box | | | |
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .923 | .921 | .922 | .001 | .721 | .722 | .734 | .011 | .923 | .922 | .923 | .001 | .721 | .717 | .730 | .013 |
| eeg | .739 | .742 | .745 | .003 | .908 | .882 | .906 | .024 | .739 | .735 | .740 | .005 | .908 | .892 | .920 | .028 |
| magic | .883 | .877 | .880 | .003 | .798 | .800 | .818 | .018 | .883 | .880 | .883 | .003 | .798 | .787 | .811 | .024 |
| phishing | .985 | .983 | .985 | .002 | .907 | .890 | .920 | .030 | .985 | .982 | .986 | .004 | .907 | .887 | .920 | .033 |
| phoneme | .888 | .878 | .885 | .007 | .783 | .774 | .814 | .040 | .887 | .880 | .889 | .009 | .782 | .755 | .809 | .054 |
| ailerons | .919 | .906 | .909 | .003 | .771 | .789 | .812 | .023 | .919 | .914 | .919 | .005 | .771 | .761 | .791 | .030 |

Table A15: Interval-valued component-conditional accuracy estimates for 70% transparency. Accuracies not covered by the corresponding Venn interval are underlined.

| | Marg. Venn | | | | | | | | Cond. Venn | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White-box | | | | Black-box | | | | White-box | | | | Black-box | | | |
| Dataset | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size | Acc. | Lower | Upper | Size |
| adult | .868 | .867 | .868 | .001 | .707 | .713 | .732 | .019 | .868 | .867 | .868 | .001 | .707 | .702 | .727 | .025 |
| eeg | .706 | .706 | .710 | .004 | .913 | .864 | .905 | .041 | .706 | .704 | .708 | .004 | .911 | .873 | .930 | .057 |
| magic | .843 | .840 | .843 | .003 | .770 | .773 | .804 | .031 | .843 | .842 | .845 | .003 | .768 | .743 | .791 | .048 |
| phishing | .951 | .948 | .952 | .004 | .891 | .862 | .912 | .050 | .951 | .947 | .951 | .004 | .891 | .852 | .920 | .068 |
| phoneme | .820 | .815 | .823 | .008 | .779 | .758 | .818 | .060 | .820 | .816 | .825 | .009 | .770 | .719 | .824 | .105 |
| ailerons | .871 | .863 | .867 | .004 | .748 | .786 | .820 | .034 | .871 | .867 | .872 | .005 | .746 | .730 | .789 | .059 |

Table A16: Interval-valued component-conditional accuracy estimates for 90% transparency. Accuracies not covered by the corresponding Venn interval are underlined.