Artificial Intelligence In Medicine 174 (2026) 103366

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Calibration-informed metrics for instance-level predictive reliability in medical AI

Federico Cabitza *

Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy
Digital Health & Wellbeing Center, Fondazione Bruno Kessler (FBK), Via Sommarive, 18, Povo, Trento, 38123, Italy

ARTICLE INFO

Keywords:
Calibration
Confidence interval
Credible intervals
Medical ML
Transparent AI

ABSTRACT

Conventional performance metrics in clinical decision support systems, such as accuracy or sensitivity, fail to reflect the reliability of individual predictions—an essential concern for clinicians operating in high-stakes environments.

We introduce a calibration-informed framework featuring two novel metrics: the Local Predictive Value (LPV) and the Credible Predictive Value (CPV). LPV estimates the empirical reliability of a prediction by assessing the observed correctness frequency in the neighborhood of its confidence score. CPV refines this estimate using a Bayesian approach, integrating global predictive values as priors to produce a posterior distribution over correctness probabilities. LPV offers a descriptive, data-driven view of local reliability, while CPV provides a belief-adjusted estimate that mitigates overfitting to sparse local data.

Applied to benchmark medical imaging datasets, these metrics yielded locally adaptive, interpretable reliability estimates. Divergences between LPV and CPV identified cases where local evidence was insufficient or misleading, highlighting how Bayesian smoothing improves stability against sparse or misleading local evidence.

By combining local calibration with Bayesian inference, LPV and CPV advance the development of medical AI systems that are not only accurate but also interpretable and trustworthy at the individual case level.

1. Introduction

Trustworthiness and transparency are widely recognized as fundamental requirements in the main legislative frameworks governing the use of AI systems in critical contexts such as medicine [1]. However, current frameworks, such as the European regulations, risk reducing compliance to merely fulfilling procedural and reporting obligations [2,3]. In practice, this often reduces compliance to the manufacturer's obligation to provide a set of documents describing the nominal and generic capabilities of these models, usually evaluated under laboratory conditions.

Clinicians should not rely solely on aggregate performance metrics, particularly when derived from datasets with complex or uncertain representativeness regarding real clinical cases [4]. For this reason, practitioners or deployers must be able to understand how, and to what extent, a model's recommendations are reliable, that is, truly trustworthy.

An initial recommendation is that developers should systematically incorporate confidence scores into their models' outputs, i.e., estimates of the probability that each prediction is correct. For example, in [5], the authors found that presenting users with alternative classifications along with their associated confidence scores generally increased both users' confidence in their own decisions and their perceived usefulness of the AI system. Similarly, the study in [6] demonstrated that providing users with confidence indicators improved human–AI collaboration. It showed that when users could appropriately calibrate their trust in the AI system based on its confidence levels, decision-making outcomes improved and both over-reliance and under-reliance on the system were mitigated.

However, case-specific scores generated by AI systems can themselves be unreliable, or even potentially misleading, if users were not also informed about the model's calibration [7,8], that is, the extent to which its probability estimates accurately reflect the actual frequency of correct predictions [9].

This leads to another key recommendation: manufacturers should provide, alongside the metrics already commonly included in model cards and technical specifications such as precision or positive predictive value, PPV, and negative predictive value, NPV, at least one calibration measure. Ideally, calibration should be reported not only at

* Correspondence to: Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy.
E-mail address: federico.cabitza@unimib.it.

https://doi.org/10.1016/j.artmed.2026.103366
Received 3 October 2025; Received in revised form 17 December 2025; Accepted 20 January 2026
Available online 5 February 2026

the global level but also across relevant ranges of confidence scores, i.e., calibration bins.

As we will show in this work, knowing both the model's confidence score for each prediction and its calibration level is essential for assessing the system's trustworthiness and for determining how much weight to place on its specific recommendations. However, additional measures are still needed.

In specialist literature, this objective is to estimate the extent to which a single probabilistic prediction approximates the so-called 'true correctness likelihood' [10], a task often denoted as "per-instance predictive uncertainty estimation" [11,12]. This latter task relates to quantifying, for a given prediction on a specific input x, how reliable the model's confidence score is, i.e., how likely the predicted label (or regression output) is to be correct. This is different from model calibration: while this latter task aims to adjust probabilities so that, on average across many instances, predicted confidence matches empirical correctness likelihood, per-instance uncertainty estimation seeks to provide an uncertainty measure tailored to each prediction, which may account for epistemic uncertainty (model uncertainty due to limited data or parameters), aleatoric uncertainty (irreducible noise in the data), or distributional uncertainty (inputs lying outside the training distribution).

1.1. Motivation and gap: what physicians are told and what they need

The performance of binary classifiers is often presented in terms of accuracy [13]: the proportion of correct suggestions out of the total provided, or equivalently, one minus the error rate. While accuracy is a familiar and intuitive metric, it is inherently a marginal measure, as all other metrics through which AI models (classifiers) are evaluated, such as sensitivity, specificity, and their means: they all reflect average performance over a population of cases (and a range of decision thresholds, as in the case of the AUC score), but provide limited insight into the reliability of individual predictions [14,15].

In requirement-gathering interviews with potential physician users, a recurrent concern emerged [16,17]: They were surprised, and often frustrated, that the systems did not provide what they perceived as the most useful piece of information for acting in the patient's best interest: the probability that a particular suggestion is correct [18].

Several responses have been proposed, but these are inadequate, particularly in high-stakes domains such as medicine, where each decision can have legally, economically, and ethically significant consequences.

A first common approach is that the model itself provides a confidence score, often referred to as the predicted probability associated with its output. A second response points to the model's predictive performance on a validation dataset, specifically the positive predictive value (PPV) for positive predictions and the negative predictive value (NPV) for negative ones.

However, the assumptions underlying these practices are weak and frequently violated in complex, real-world settings. The response about the predictive value, in particular, rests on a problematic assumption [19]: that a marginal performance metric, i.e., one computed over a sample—can be reliably generalized to every individual instance, or even to any other instance drawn from the same population (or from a similarly distributed one). On the other hand, the response about confidence assumes that the model is well-calibrated, that is, capable of accurately estimating the probability that each of its predictions is correct. Yet, this is rarely the case, especially with many of the most effective and widely adopted modern models [8,10]. If calibration were reliably achieved, the need for additional methods of uncertainty quantification like what we will propose in what follows would be substantially diminished.

In this work, we introduce a metric framework that allows even non-expert decision-makers to make use of marginal information — such as a model's average predictive performance (e.g., PPV and NPV) — in conjunction with the model's specific output, and calibration-related evaluation to assess the reliability of individual classifications. This enables decision-makers to make informed judgments at the instance level, offering actionable insights into the reliability of each specific prediction.

Although this is precisely what clinicians and other users of decision support systems require [18,20], and despite the growing emphasis on model transparency and interpretability, our framework addresses a notable research gap: there remains no widely adopted metric that provides case-by-case reliability estimates in an actionable and practically meaningful way [7,21].

In the following sections, we present this metric framework, describe its implementation, and apply it to three widely used clinical benchmarks to illustrate both its interpretability and its potential to enhance clinicians' understanding of the reliability of outputs provided by ML- or AI-based classification systems. In the discussion, we also address the broader implications of our proposal (or similar ones) for transparency and regulatory compliance. In particular, we highlight practices that would encourage providers to publish test data in full—including true labels and model-predicted probabilities-on external test sets, alongside traditional marginal metrics such as those derived from the confusion matrix.

1.2. Main contribution

This work addresses the lack of per-instance reliability estimates that jointly account for local model calibration and prior predictive performance. To this end, we propose a Calibration-Informed Predictive Value (CIPV) framework that transforms model confidences and observed outcomes into two complementary quantities: a Local Predictive Value (LPV), defined in a frequentist way from local empirical evidence, and a Credible Predictive Value (CPV), obtained through a Bayesian update of prior predictive values. We illustrate the behavior and interpretability of this framework in three case studies based on clinical imaging data. Our main contribution is a model-agnostic and clinically interpretable reliability layer that can be attached to any probabilistic classifier, providing per-instance predictive values and uncertainty intervals that are aligned with familiar measures such as positive and negative predictive values.

From a methodological standpoint, our framework complements existing local calibration diagnostics and per-instance uncertainty estimation methods. Prior work on calibration has largely focused on aggregate or bin-level measures (e.g., reliability diagrams, global or local ECI scores), which do not directly yield per-instance predictive values with uncertainty intervals that clinicians can interpret as variants of PPV or NPV. Conversely, many per-instance uncertainty scores, such as those based on deep ensembles, Bayesian approximations, or post-hoc confidence calibration, are often model-specific and less directly interpretable to end users. In contrast, the proposed framework integrates (i) a neighborhood-based local calibration index, (ii) global priors such as PPV/NPV or disease prevalence, and (iii) a dual frequentist-Bayesian quantification (LPV and CPV with uncertainty intervals) into a single, model-agnostic reliability layer that can be attached to any probabilistic classifier.

2. Calibration-informed reliability framework

Our framework integrates three complementary information sources: (i) the model's own confidence score for the predicted class, (ii) the degree of empirical calibration in the neighborhood (bin) of that score—quantified using a SOTA calibration metric, the local Estimated Calibration Index (local ECI [9]), and (iii) the model's marginal Positive Predictive Value (PPV) and Negative Predictive Value (NPV) as global priors. The resulting measures yield a probability of correctness that is locally adaptive, globally informed, and directly interpretable by clinicians in their decision-making process.

Fig. 1 provides an overview of the architecture and computational flow of the proposed calibration-informed reliability framework.

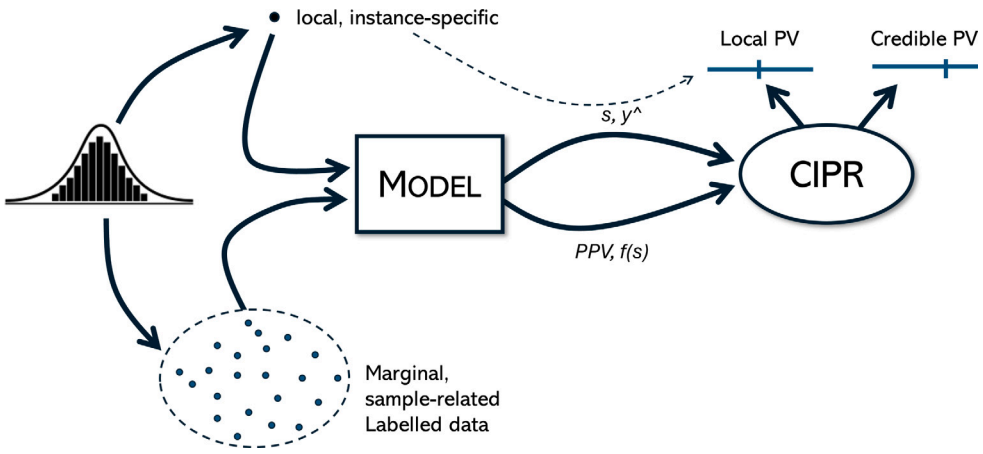


Fig. 1. Architecture and computational flow diagram of the calibration-informed reliability framework. The diagram illustrates the estimation of Local Predictive Value (LPV) and Credible Predictive Value (CPV) from model outputs, calibration data, and local neighborhoods.

2.1. Problem formulation and notation

Let X_i denote the feature vector of instance i. For each i, the classifier outputs a probability estimate $\hat{p}_i \in [0, 1]$ for the positive class ($Y = 1$). The predicted label is then defined as $\hat{y}_i \in \{0, 1\}$, obtained by thresholding $\hat{p}_i$. We define the associated confidence score as

$$s_i := P(Y = \hat{y}_i \mid X_i),$$

as estimated by the model; that is,

$$s_i = \hat{p}_i \quad \text{if } \hat{y}_i = 1, \qquad s_i = 1 - \hat{p}_i \quad \text{if } \hat{y}_i = 0.$$

Also, the positive predictive value (PPV) and the negative predictive value (NPV) of the classifier evaluated on a set of instances are:

$$\text{PPV} = P(Y = 1 \mid \hat{y} = 1), \qquad \text{NPV} = P(Y = 0 \mid \hat{y} = 0).$$

In contrast, the calibration of the classifier concerns the agreement between its predicted confidence scores and true outcome frequencies; therefore the model is said to be perfectly calibrated if, for every probability level $p \in [0, 1]$,

$$P(Y = 1 \mid s_i = p) = p.$$

That is, among all instances assigned a confidence score p, the empirical proportion of positives equals p.

That said, we seek a local, instance-specific measure that answers the question:

> "What is the probability that this particular prediction is correct, given the model's confidence and its empirical calibration at that confidence level?"

We denote this quantity as a *Calibration-Informed Predictive Value* (CIPV), that is an estimation of:

$$P(Y = \hat{y}_i \mid s_i),$$

that is, the conditional probability that the predicted label matches the true outcome given the confidence score s_i.

Operationalization of the estimator. We propose two complementary CIPVs, depending on how local calibration and priors are combined:

1. a locally calibrated point estimate with its calibration-based reliability interval, which we call *Local Predictive Value*, or LPV;
2. a Bayesian posterior mean with its credible interval, which we call *Credible Predictive Value*, or CPV;

Consequently, we describe the computation of:

1. the **Local Predictive Value, LPV.** The model's raw confidence score is adjusted based on the empirical event rate observed in a neighborhood of similar cases. For example, if the model outputs a 70% probability of disease and the local observed frequency is 65%, the calibrated estimate might be shifted downward to 68%. This provides a point estimate that reflects local calibration and may be more conservative than the unadjusted model score. The LPV comes with a **LPV reliability interval**, that is an uncertainty band around the point estimate, derived solely from empirical variation in the local calibration neighborhood. Continuing the previous example, this might yield an interval such as [0.60, 0.80], indicating that — based on historical cases with similar predicted confidence — the true probability plausibly lies within this range. Narrow intervals suggest high local consistency; wider intervals indicate greater calibration uncertainty.
2. the **Credible Predictive Value, CPV.** This is the expected value (mean) of a Bayesian posterior distribution (e.g., 67%), accompanied by its credible interval (e.g., [0.61, 0.72] at the 95% level), obtained from the quantiles of that same posterior distribution to quantify residual uncertainty. By Bayesian posterior distribution of the model's predictive performance, we mean the range of plausible values that the model's predictive value can take, after combining the evidence from the observed predictions and outcomes with prior assumptions. Instead of reporting only a single number, the posterior distribution allows us to give both a best estimate and a credible interval that quantifies the uncertainty around that estimate. In other words, the CPV is a fully probabilistic summary that integrates global prior knowledge with local calibration evidence.

Further details on the practical interpretation of these two estimators are provided in Section 3.1.

Neighborhood selection. Local calibration and predictive value estimates are based on a reference neighborhood defined around s_i in the confidence space, also denoted as a *bin*.

Two modes are supported. By default, local calibration is computed using a *k-nearest neighbors* (k-NN) approach in confidence space. Given a reference score s_i, the k closest instances are identified (typically $k = 50$, adjustable by the user). Each neighbor contributes either equally or with a kernel weight that decays with distance from s_i according to alternative logic (uniform, triangular, Gaussian) modulating each neighbor's contribution based on distance from s_i. This procedure ensures that a sufficient and fixed number of observations is consistently considered, while reducing bias through distance-sensitive weighting.

As an alternative, the framework also supports an *adaptive binning* mode. In this case, a confidence interval is constructed around s_i with an initial default half-width ($\Delta = 0.05$). If s_i lies near the boundaries 0 or 1, the interval expands inward (e.g., $[0, s_i + 2\Delta]$). If the bin contains fewer than a certain amount of instances (e.g., $n_{\min} = 50$), its half-width is expanded geometrically (factor 1.5) up to a maximum of 0.25. Conversely, if the bin becomes too large (exceeding, e.g., $n_{\max}$, e.g. 200), it is shrunk multiplicatively (with a default factor of 0.8) down to the base margin.

In both neighborhood modes, an optional *scope* restriction allows the computation to include only reference instances that share the same predicted label $\hat{y}_i$ as the case under evaluation. This yields an estimate that is more indicative and clinically relevant, because calibration is assessed within the same decision class (e.g., only positive predictions). However, this restriction reduces the number of available samples and may increase statistical uncertainty. Thus, researchers and practitioners must balance two competing aspects: the representativeness of LPV and CPV at the class level, and the reliability of the estimates given the available sample size.

Geometric definition of local calibration. Within the final bin $\mathcal{B}$, let

$$x_b = \frac{1}{n_b} \sum_{j \in \mathcal{B}} s_j, \qquad y_b = \frac{1}{n_b} \sum_{j \in \mathcal{B}} y_j$$

denote the average model confidence and the empirical positive event rate, respectively in the neighborhood. Although [9] defines local calibration using the *L2* (perpendicular) distance to the identity (diagonal) line, here we adopt the *L1* (Manhattan/vertical) formulation, which is more intuitive on reliability diagrams—while yielding the *same numerical index* as L2 when the normalization is applied consistently (see Fig. 2).

We therefore measure the local deviation as the shortest, vertical gap between the point (x_b, y_b) and the identity diagonal line $y = x$ of perfect calibration, that is:

$$d \;=\; |\, y_b - x_b \,|,$$

and normalize it by the worst-case vertical deviation at x_b,

$$d_{\max}(x_b) \;=\; \max\{\, x_b,\; 1 - x_b \,\}.$$

The local calibration index is then

$$\mathrm{ECI}_{\mathrm{local}} \;=\; 1 - \frac{d}{d_{\max}(x_b)} \;\in\; [0, 1].$$

Thus, $\mathrm{ECI}_{\mathrm{local}} = 1$ indicates perfect local calibration ($y_b = x_b$), whereas $\mathrm{ECI}_{\mathrm{local}} = 0$ corresponds to maximal miscalibration at that confidence level. Additionally, we record whether the model is, in that bin, *over-confident* ($y_b < x_b$), *under-confident* ($y_b > x_b$), or *well-calibrated* ($y_b \approx x_b$).

The following sections introduce the formal definitions of LPV and CPV, supplemented by intuitive explanations and illustrative examples in Section 4, which show how these metrics behave on real-world clinical cases. Readers primarily interested in interpretation and usage may wish to consult Sections 3.1 and 4.4 before returning to the full technical details.

2.2. Local predictive value (LPV)

Building on $\mathrm{ECI}_{\mathrm{local}}$, we define the *Local Predictive Value (LPV)* a predictive value s_i, adjusted toward the empirical event rate y_b, that accounts for the reliability of calibration around s_i. The point estimate of the LPV is computed through the empirical event rate in the bin as follows:

$$\mathrm{LPV} = \begin{cases} \mathrm{ECI}_{\mathrm{local}} \cdot s + (1 - \mathrm{ECI}_{\mathrm{local}}) \cdot y_b & \text{if } \hat{y} = 1, \\ \mathrm{ECI}_{\mathrm{local}} \cdot s + (1 - \mathrm{ECI}_{\mathrm{local}}) \cdot (1 - y_b) & \text{if } \hat{y} = 0. \end{cases}$$

with $\hat{y} \in \{0, 1\}$ the model's predicted label and s the associated confidence score.

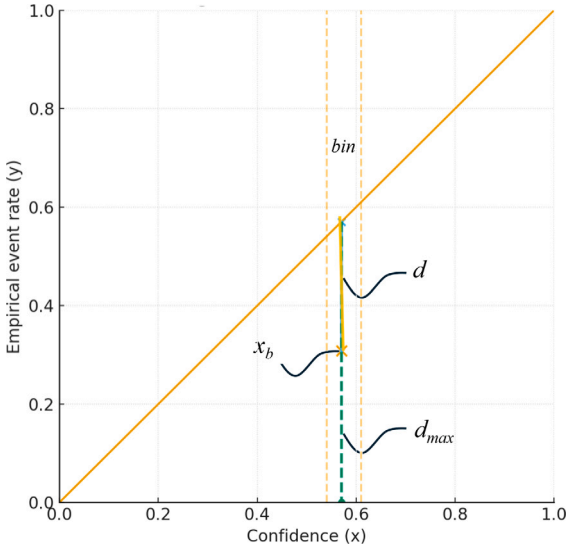


Fig. 2. Illustration of local calibration error $d = |y_b - x_b|$, defined as the vertical (L1) distance from the average bin point (x_b, y_b) to the identity line $y = x$, normalized by the worst-case deviation $d_{\max}(x_b) = \max\{x_b, 1 - x_b\}$.

Intuitively, LPV adjusts the raw confidence score based on the frequency of correct outcomes in similar past predictions. For instance, if the model predicts a disease with 80% confidence, but in similar past cases the outcome was correct only 65% of the time, the LPV may yield a more conservative reliability estimate closer to 70%. This reflects the model's empirical trustworthiness in the local region of the score space.

To quantify uncertainty, a reliability interval I_{ECI} is derived from a radius ϵ that quantifies the uncertainty attributable to local miscalibration. This radius is defined using the mean confidence within the bin, x_b, as follows:

$$\epsilon = (1 - \mathrm{ECI}_{\mathrm{local}}) \cdot \max(x_b, 1 - x_b),$$

The resulting reliability interval, I_{ECI}, is centered on the instance's confidence score s and is given by:

$$I_{\mathrm{ECI}} = [\max(0, s - \epsilon), \min(1, s + \epsilon)]$$

This formulation ensures that the interval remains bounded within $[0, 1]$.

2.3. Credible predictive value (CPV)

To capture statistical uncertainty, we also consider confidence scores within a Bayesian framework. In the case of the Credible predictive value (CPV) the predictive value is modeled as the posterior mean of a Beta distribution, with prior counts (A_0, B_0) derived either from global dataset statistics, such as PPV and NPV, or from a pooled prior that combines global and local bin frequencies with a shrinkage factor κ.

In the Bayesian formulation, the prior distribution is set to reflect the model's global predictive value. Specifically, when the predicted label is positive ($\hat{y} = 1$), the prior hyperparameters are derived from the global counts of true positives and false positives, so that the prior mean equals the global PPV. Conversely, when the predicted label is negative ($\hat{y} = 0$), the prior uses the counts of true negatives and false negatives, ensuring that the prior mean equals the global NPV.

Calibration quality is translated into pseudo-counts that update the Beta posterior: the number of virtual successes is proportional to the prediction score, and the number of virtual failures to its complement, both scaled by a weight that reflects local calibration. We support both an evidence-based mapping and an uncertainty-based mapping, with

a scaling λ that can be constant or proportional to the bin size. By default, the implementation enables a pooled prior. If Beta quantiles are unavailable, credible interval endpoints are left undefined.

Formally, the posterior parameters (α, β) are computed as:

$$\alpha = A_0 + w \cdot s, \qquad \beta = B_0 + w \cdot (1 - s)$$

The Credible Predictive Value (CPV) is then defined as the posterior mean of the Beta distribution:

$$\text{CPV} = \frac{\alpha}{\alpha + \beta}$$

From this posterior, we also compute a credible interval at a specified level, capturing the residual epistemic uncertainty around the prediction's correctness.

In intuitive terms, the CPV expresses our degree of belief that the model's prediction is correct, after combining the local confidence score and calibration quality with global predictive values (PPV or NPV) as priors. When local evidence is strong and well-calibrated, the CPV aligns closely with the LPV. However, when the local neighborhood is sparse or unreliable, the CPV leans more heavily on the prior, offering a tempered estimate. The credible interval accompanying the CPV reflects this uncertainty: narrower intervals suggest strong, consistent evidence, while wider intervals indicate limited or ambiguous support for the prediction.

ECI-to-weight mappings and scale selection. The instance weight w, which translates local calibration quality into pseudo-counts, is determined by a scaling factor for pseudo-counts λ, controlling how much statistical weight to give to a locally calibrated score; and a mapping function. Two mappings are implemented:

- **Evidence-based mapping:** This mapping is prudential, down-weighting instances from small bins, even if calibration seems good. It is stabilized by a constant κ_{evidence}, which controls trust in small samples acting like a "soft prior size" or pseudo-count comparator (default 30.0):

$$w = \lambda \cdot \text{ECI}_{\text{local}} \cdot \left(\frac{n_{\text{bin}}}{n_{\text{bin}} + \kappa_{\text{evidence}}} \right)$$

- **Uncertainty-based mapping:** This mapping assigns higher weight to instances where the calibration-based reliability interval is narrow. A small stabilizer ϵ_{uncert} (default 10^{-6}) prevents division by zero:

$$w = \lambda \cdot \frac{\text{ECI}_{\text{local}}}{(1 - \text{ECI}_{\text{local}}) \cdot \max(s, 1 - s) + \epsilon_{\text{uncert}}}$$

The scale factor λ can be set as a constant (e.g., 30) or adaptively as $\lambda = \max(10, 0.1 \cdot n_{\text{bin}})$, so that local data are trusted proportionally to their number, which is consistent with Bayesian reasoning.

Pooled prior. When local bin counts are available, a pooled prior is constructed. The blending factor, $\rho \in [0, 1]$, modulates the weight of local counts based on bin size n_{loc}, local calibration ECI_{local}, and a shrinkage parameter κ_{pool}:

$$\rho = \frac{n_{\text{loc}}}{n_{\text{loc}} + \kappa_{\text{pool}}} \cdot \text{ECI}_{\text{local}}.$$

In other words, ρ can be interpreted as a "trust coefficient" for local data, which controls the relative weight between local evidence and global priors when setting the hyperparameters (A_0, B_0) of the Beta prior for Bayesian inference, thus affecting the Bayesian posterior mean and the credible interval to stabilize inference, and avoid overfitting to unreliable local data.

Let $(C_1^{\text{glob}}, C_2^{\text{glob}})$ be the global counts (e.g., TP, FP) and $(C_1^{\text{loc}}, C_2^{\text{loc}})$ be the local counts. The hyperparameters of the pooled Beta prior, (A_0, B_0), are formed by adding these counts, weighted by ρ, to a base prior (e.g., a uniform prior with $\alpha_{\text{prior}} = 1, \beta_{\text{prior}} = 1$):

$$A_0 = \alpha_{\text{prior}} + (1 - \rho) \cdot C_1^{\text{glob}} + \rho \cdot C_1^{\text{loc}}, \quad B_0 = \beta_{\text{prior}} + (1 - \rho) \cdot C_2^{\text{glob}} + \rho \cdot C_2^{\text{loc}}.$$

Robustness and branch conditions. If not provided, the bin size is inferred from local confusion counts. The Bayesian branch is computed only if the relevant global counts are available. In cases where ECI_{local} cannot be computed (e.g., if the bin has fewer instances than the required minimum), its value defaults to 0.0. This treats instances in unreliably small neighborhoods as maximally uncalibrated, causing the Bayesian estimator to rely entirely on the prior.

2.4. Summary: Formal definitions of LPV and CPV

LPV definition. The *Local Predictive Value (LPV)* is a reliability-adjusted score computed from the model's raw confidence s and the local empirical outcome rate y_b, weighted by the local Estimated Calibration Index ($\text{ECI}_{\text{local}}$):

$$\text{LPV} = \begin{cases} \text{ECI}_{\text{local}} \cdot s + (1 - \text{ECI}_{\text{local}}) \cdot y_b & \text{if } \hat{y} = 1 \\ \text{ECI}_{\text{local}} \cdot s + (1 - \text{ECI}_{\text{local}}) \cdot (1 - y_b) & \text{if } \hat{y} = 0 \end{cases}$$

This estimate reflects local calibration and observed reliability in the neighborhood of similar predictions.

CPV definition. The *Credible Predictive Value (CPV)* is the posterior mean of a Beta distribution constructed from global prior counts and local calibration-informed pseudo-counts:

$$\text{CPV} = \frac{\alpha}{\alpha + \beta}$$

where the parameters α and β are defined as:

$$\alpha = A_0 + w \cdot s, \quad \beta = B_0 + w \cdot (1 - s)$$

The CPV balances local evidence with global priors to yield a posterior estimate of prediction correctness, accompanied by a credible interval.

3. Interpretability and practical use

3.1. Interpretation by non-specialists

The metrics defined within our calibration-informed framework are statistical constructs, but their interpretation can be communicated in terms accessible to clinicians and other non-specialists. They address a straightforward yet critical question: *"For this individual case, how likely is it that the model's prediction is correct?"* Unlike conventional predictive values that reflect only overall model accuracy, LPV and CPV incorporate two complementary sources of evidence (see Fig. 1) as priors: (i) the model's historical performance on a large test or validation set, and (ii) its calibration in the neighborhood of cases with similar confidence scores.

From a clinical standpoint, this layered interpretation serves distinct needs and can be framed as two related but separate sub-questions.

The LPV helps clinicians address questions such as: "How should we adjust the model's confidence score based on its typical tendency to overestimate or underestimate its predictive accuracy for similar cases?" By contrast, the CPV supports a broader inquiry: "Given the model's confidence in its prediction, its calibration, and its prior performance on similar cases, what is the probability that this specific prediction is correct?"

Accordingly, we propose the CPV point estimate as the primary actionable metric, with its associated credible interval quantifying uncertainty. This interval is analogous in form to a confidence interval but differs in interpretation: "Given the observed data (the raw confidence score) and the priors about the marginal predictive values and the calibration level in the score's bin, there is a high probability (95%) that the true predictive value, that is, the probability that the prediction is correct—lies within this range".

The LPV point estimate and its corresponding interval, in turn, provide insight into local model performance. The LPV interval is a confidence interval in the frequentist sense: it reflects the long-run

reliability of the estimation procedure. The correct interpretation is: "If the procedure that produced this interval were repeated on many independent samples from the same population, then 95% of the resulting intervals would contain the true predictive value". In this framework, one does not assign a probability to the true value lying in the specific interval at hand, but one can be highly confident in the procedure that generated it. LPV intervals are therefore useful for hypothesis testing: for instance, if the entire interval lies above a threshold (e.g., >0.5), this provides evidence to reject the hypothesis that the true predictive value is at or below that threshold.

In practical terms, narrow intervals and point estimates close to the model's raw probability suggest reliable performance. Conversely, wide intervals or substantial adjustments indicate that the raw output may be less reliable, warranting greater caution and increased reliance on clinical judgment. Additional contextual indicators (e.g., local ECI, bin size) further inform the trustworthiness of the estimates for the individual patient.

Software implementation. All estimators described above are implemented in a public Python script and a privacy-preserving web tool (see the Appendix). These tools allow practitioners to compute LPV and CPV estimates given the true labels and the confidence scores collected on a test/validation dataset. The Appendix also provides detailed documentation of the algorithmic parameters and configuration options.

4. Case studies on clinical datasets

For illustrative purposes, we applied the Calibration-Informed framework to models developed within three popular clinical benchmarks: *PathMNIST*, *DermaMNIST*, and *PneumoniaMNIST*. The goal was to show the range of values produced in real-world datasets and how they can be interpreted under varying conditions of empirical support and local calibration. For each model and dataset, we selected representative test cases and analyzed the outcomes of both LPV and CPV estimates, focusing on four key aspects:

- Local empirical frequency (i.e., the observed proportion of correct predictions in the neighborhood).
- Local calibration (quantified by the Estimated Calibration Index, ECI).
- Width of the confidence bin, which reflects the reliability of local information.
- Coherence between local evidence, the global prior, and the posterior update.

We selected three models representative of currently deployed medical diagnostic systems and that exhibit high discriminative capacity, with an AUC above 90% (see Table 2). However, one of them, in the pneumaMNIST benchmark, shows relatively low accuracy, close to the threshold of clinical utility. This model achieves very high sensitivity — higher than that of the more accurate models — but only modest specificity, marginally better than that obtained by randomly labeling cases as negative. By contrast, the other two models achieve excellent specificity and at least acceptable sensitivity, resulting in high negative predictive values (NPV).

The models also differ substantially in calibration (see Table 3), which is particularly important for the metrics defined in our framework. The first two models, characterized by high accuracy and specificity, exhibit calibration errors below 5% and 2% (see the ECE score). In contrast, the high-sensitivity model is poorly calibrated, with an error approaching one-third of the maximum possible. This is reflected in a very high Brier score (more than twice that of the second model) and a reliance component nearly twenty times larger.

Another important distinction lies in the evaluation datasets. The first two specific models were tested on large datasets of 7000 and 2000 cases, respectively, whereas the high-sensitivity model was tested on only about 600 cases, which makes local calibration assessment particularly challenging.

Thus, although all three models demonstrate strong discriminative performance, they differ widely in accuracy, sensitivity–specificity trade-offs, calibration quality, and test set size. This diversity provides a rich basis for illustrating the application and relevance of the LPV and CPV scores. All subsequent cases have been selected for illustrative purposes only and are summarized in Table 1.

4.1. Case study 1: PathMNIST

The PathMNIST test dataset was designed to benchmark models for predicting survival from colorectal cancer histology slides [22]. It includes a test set of 7180 image patches collected from a clinical center distinct from those that supplied the 100,000 non-overlapping hematoxylin-and-eosin-stained histological image patches used for training and validation. The dataset originally contains nine tissue classes, which we consolidated into two categories — normal tissue versus cancerous or pre-cancerous tissue — thereby framing the task as binary classification. We considered the predictive performance of the Google AutoML Vision model, that on the original test dataset achieved an AUC of 0.941 and an accuracy of 0.744 [23]. Its performance on the binarized test dataset is reported in Tables 2 and 3, while its calibration plot is depicted in Fig. 3.

Case 1: Positive prediction with low confidence. The model produced a positive prediction with low confidence. The local neighborhood contained a nearly balanced mixture of positives and negatives (47 vs. 53, $y_{\text{mean}} = 0.45$), resulting in a modest LPV estimate of 0.53 and a wide reliability interval [0.45, 0.65]. Local calibration was nevertheless high ($\text{ECI}_{\text{local}} = 0.81$), showing that scores in this region were internally consistent. The CPV rose to 0.76 with a tighter credible interval [0.72, 0.79], demonstrating how the Bayesian update stabilizes local evidence by integrating global prevalence (see Fig. 4).

Case 2: Negative prediction with high confidence. Here the classifier returned a negative prediction with high confidence (since a low probability for the positive class implies strong support for the negative class). The bin contained 74 true negatives and 26 false negatives ($y_{\text{mean}} = 0.32$), yielding an LPV of 0.82 with a broad interval [0.68, 1.00]. Although local calibration was moderate ($\text{ECI}_{\text{local}} = 0.79$), the positive rate was not negligible. The CPV increased to 0.91 with a narrow interval [0.90, 0.92], reflecting the stabilizing effect of global evidence: across the dataset, negative predictions with low y_{prob} are typically reliable. This case illustrates how CPV leverages prior knowledge to reinforce reliability where the local estimate alone is uncertain.

Case 3: Positive prediction with very high confidence. The model produced a very high-confidence positive prediction. Local evidence was overwhelmingly strong, with 98 positives and only 2 negatives ($y_{\text{mean}} = 0.96$), and calibration was nearly perfect ($\text{ECI}_{\text{local}} = 0.97$). Accordingly, the LPV estimate was 0.93 with a narrow interval [0.90, 0.96]. Surprisingly, the CPV mean was lower, at 0.84 ([0.80, 0.87]). This downward adjustment reflects the conservative pooling strategy: even with strong local evidence, the posterior is moderated by a globally lower PPV. This reflects an intentional safeguard against overconfidence in small local clusters.

4.2. Case study 2: DermaMNIST

The DermaMNIST test set is derived from HAM10000, a large collection of multi-source dermatoscopic images of common pigmented skin lesions [24]. From this dataset, 2005 dermatoscopic images were extracted and categorized into seven disease classes. For our purposes, we binarized these classes into benign versus non-benign moles. We then selected the Google AutoML Vision model, which achieved an

Table 1
LPV and CPV scores, with corresponding intervals, for different cases. For each instance, a local neighborhood of 100 predictions closest in confidence to the value in column *Conf* was selected (using k-NN with triangular kernel). Within this neighborhood, x_{mean} is the average confidence score, y_{mean} the observed outcome rate, and *ECI local* the local Expected Calibration Index $(1-|y_{mean}-x_{mean}|/d_{max})$, where higher values indicate better local calibration. LPV (Local Predictive Value) intervals are reliability intervals based on ECI, while CPV (Credible Predictive Value) intervals are Bayesian credible intervals.

Benchmark	Pred	Conf	Bin interval	x_{mean}	y_{mean}	ECI local	LPV	CPV
dermaMNIST	0	0.47	(0.29, 0.64)	0.45	0.28	0.70	0.59 [0.37, 0.70]	0.96 [0.94, 0.97]]
dermaMNIST	1	0.82	(0.70, 0.94)	0.83	0.60	0.73	0.76 [0.59, 1.00]	0.65 [0.57, 0.71]]
dermaMNIST	0	0.07	(0.06, 0.08)	0.07	0.05	0.98	0.93 [0.91, 0.95]	0.96 [0.94, 0.98]]
dermaMNIST	1	0.94	(0.89, 0.99)	0.95	0.86	0.91	0.93 [0.85, 1.00]	0.76 [0.69, 0.82]]
pathMNIST	1	0.55	(0.52, 0.58)	0.55	0.45	0.81	0.53 [0.45, 0.65]	0.76 [0.72, 0.79]]
pathMNIST	0	0.14	(0.13, 0.15)	0.14	0.32	0.79	0.82 [0.68, 1.00]	0.91 [0.90, 0.92]]
pathMNIST	1	0.93	(0.93, 0.94)	0.93	0.96	0.97	0.93 [0.90, 0.96]	0.84 [0.80, 0.87]]
pneumoniaMNIST	1	0.54	(0.54, 0.54)	0.54	0.94	0.27	0.83 [0.14, 0.94]	0.79 [0.75, 0.83]]
pneumoniaMNIST	0	0.45	(0.45, 0.48)	0.47	0.00	0.13	0.94 [0.09, 1.00]	0.77 [0.71, 0.83]]

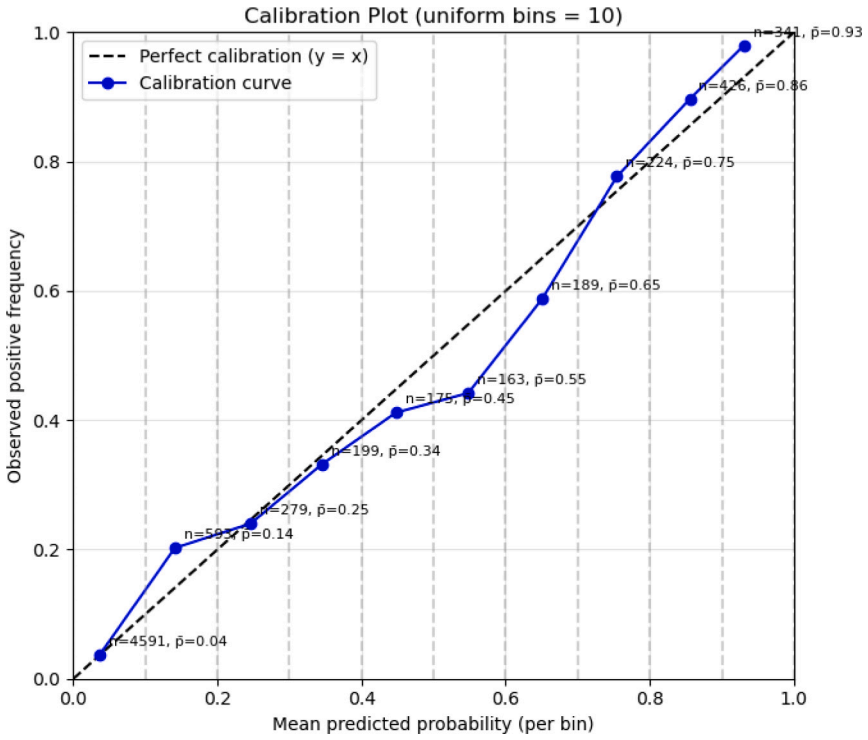


Fig. 3. Reliability diagram of the model evaluated on the pathMNIST benchmark with 10 uniform probability bins. For each bin, the annotation indicates the number of samples falling into that bin (n) and the average predicted probability within the bin ($\bar{p}$). The diagonal dashed line represents perfect calibration ($y = x$), while the blue curve shows the observed positive frequency as a function of the mean predicted probability. Vertical gray dashed lines denote bin boundaries.

Table 2
Classification and Discrimination metrics of the evaluated models (at the 0.5 decision threshold). Reported measures include accuracy (Acc.), sensitivity (Sens., true positive rate), specificity (Spec., true negative rate), marginal positive predictive value (PPV, precision), marginal negative predictive value (NPV), and area under the ROC curve (AUC).

Benchmark	Acc.	Sens.	Spec.	PPV	NPV	AUC
PathMNIST	0.893	0.681	0.952	0.799	0.914	0.919
DermaMNIST	0.928	0.727	0.950	0.613	0.970	0.947
PneumaMNIST	0.780	0.890	0.598	0.787	0.765	0.938

Table 3
Calibration metrics of the evaluated models. Reported measures include Expected Calibration Error (ECE), the Brier score, and its Murphy's decomposition into reliability (Rel.), resolution (Res.), and uncertainty (Unc.).

Benchmark	ECE	Brier	Rel.	Res.	Unc.
PathMNIST	0.017	0.080	0.001	0.091	0.171
DermaMNIST	0.047	0.047	0.005	0.047	0.089
PneumaMNIST	0.319	0.224	0.108	0.115	0.234

AUC of 0.91 and an accuracy of 0.768 on the original dataset [23]. Its performance on the binarized test dataset is reported in Tables 2 and 3, while its calibration plot is depicted in Fig. 5.

As shown in Fig. 6, divergences between LPV and CPV identify regions where local evidence is insufficient or unstable, suggesting that the CPV estimate is more robust in these settings.

Case 1: Negative prediction with low confidence ($\hat{y} = 0$, $p = 0.47$). The LPV point estimate is 0.59 with a wide interval [0.37, 0.69], reflecting substantial uncertainty in the local neighborhood. The mean observed outcome rate (28%) is well below the reference score (47%), indicating local overconfidence. In contrast, the CPV mean rises to 0.96 with a narrow credible interval [0.94, 0.97]. This divergence shows how Bayesian pooling stabilizes the posterior when global evidence strongly favors the negative class, even if local calibration is imperfect. For

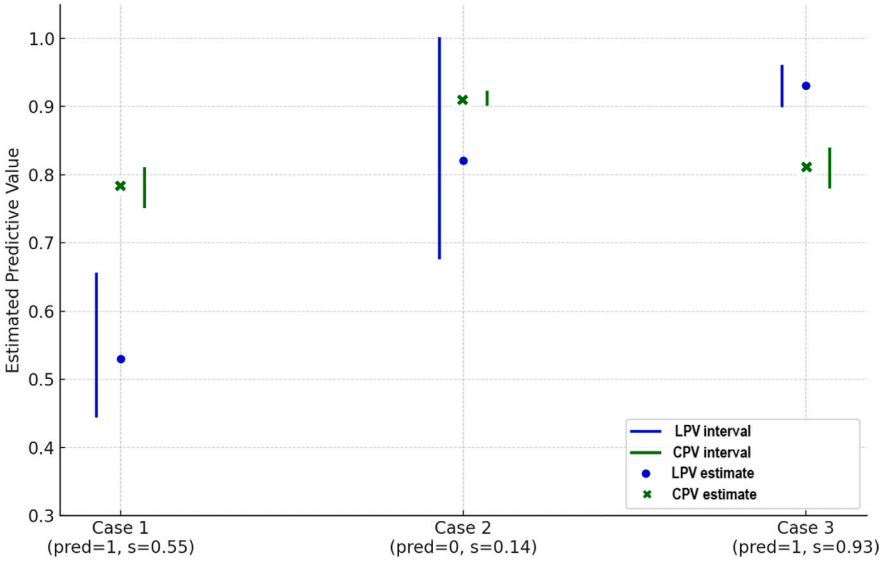


Fig. 4. Comparison of calibration-informed predictive value (CPV) and local predictive value (LPV) estimates for three representative cases in the pathMNIST benchmark. Blue markers indicate LPV point estimates with their corresponding uncertainty intervals (blue vertical lines), while green markers denote CPV point estimates with associated intervals (green vertical lines). For each case, the predicted label (*pred*) and the model's raw confidence score (s) are reported on the x-axis. The y-axis shows the estimated predictive value, highlighting differences in calibration-aware versus local empirical estimation of predictive reliability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

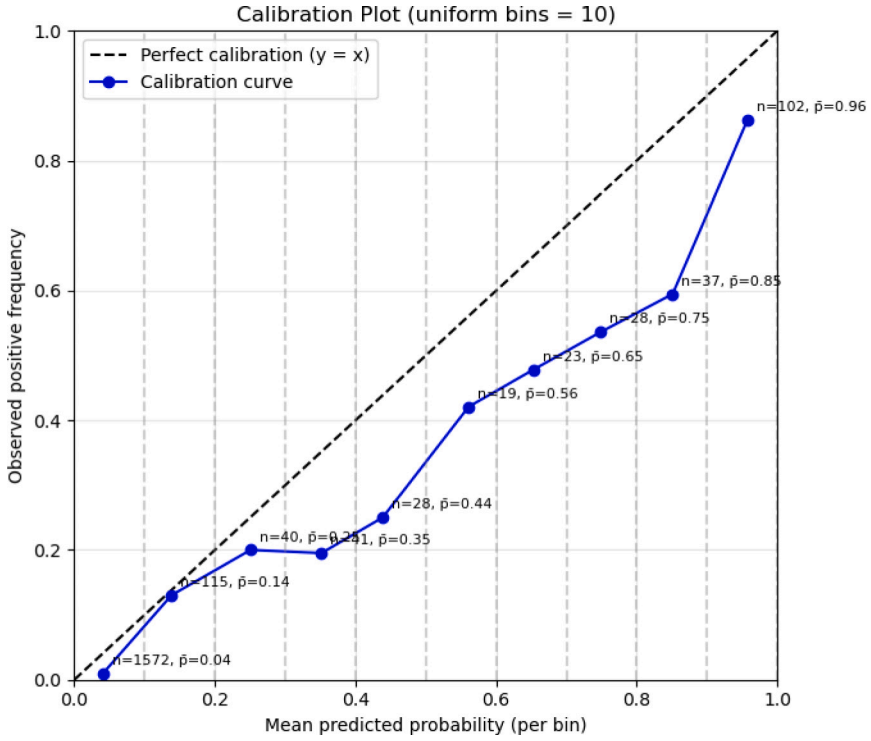


Fig. 5. Reliability diagram of the model evaluated on the dermaMNIST benchmark with 10 uniform probability bins. For each bin, the annotation indicates the number of samples falling into that bin (n) and the average predicted probability within the bin ($\bar{p}$). The diagonal dashed line represents perfect calibration ($y = x$), while the blue curve shows the observed positive frequency as a function of the mean predicted probability. Vertical gray dashed lines denote bin boundaries.

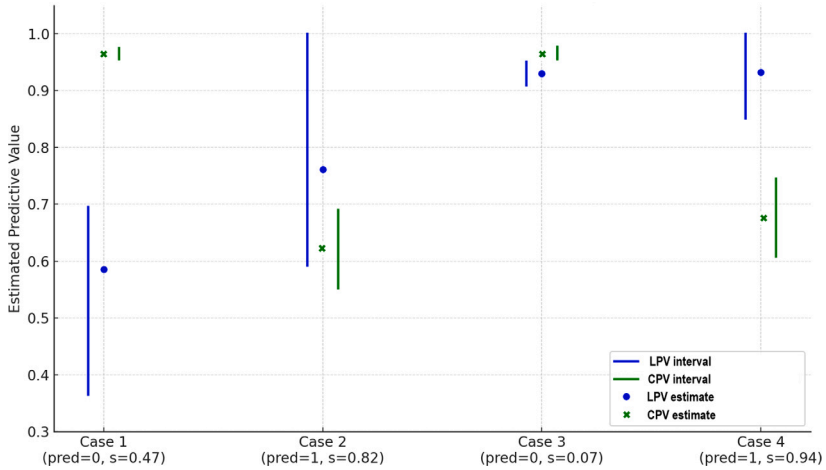


Fig. 6. Comparison of calibration-informed predictive value (CPV) and local predictive value (LPV) estimates for four representative cases in the dermaMNIST benchmark. Blue markers indicate LPV point estimates with their corresponding uncertainty intervals (blue vertical lines), while green markers denote CPV point estimates with associated intervals (green vertical lines). For each case, the predicted label (*pred*) and the model's raw confidence score (s) are reported on the x-axis. The y-axis shows the estimated predictive value, for each case, with LPV and CPV plotted on the same scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

practitioners, the message is that although the immediate neighborhood appears unreliable, the broader dataset indicates that a negative prediction at this score is indeed very trustworthy.

Case 2: Positive prediction with moderate confidence ($\hat{y} = 1$, $p = 0.82$). Here the LPV point estimate is 0.76, with an interval [0.59, 1.00]. The local calibration shows that outcomes above the decision threshold often, but not always, correspond to true positives. The CPV mean is lower, at 0.64 ([0.57, 0.71]), indicating that Bayesian pooling discounts local optimism by incorporating false positives observed in the wider dataset. This demonstrates that even apparently confident positive predictions require caution once broader calibration information is taken into account.

Case 3: Negative prediction with very high confidence ($\hat{y} = 0$, $p = 0.07$). The LPV point estimate is extremely high (0.93) with a very tight interval [0.91, 0.95]. This occurs because the reference score lies deep in the lower tail of the distribution, where almost all cases are negative. The CPV mean (0.96, [0.94, 0.98]) closely aligns with the LPV, confirming and slightly reinforcing the reliability of a negative prediction at this very low y_{prob}. This case exemplifies high-confidence predictions for the negative class, where both local and global evidence strongly agree.

Case 4: Positive prediction with very high confidence ($\hat{y} = 1$, $p = 0.94$). The LPV point estimate is again high (0.93, interval [0.85, 1.00]), consistent with the strong score. However, the CPV mean drops substantially to 0.76 ([0.69, 0.82]). This discrepancy reflects that even among highly confident positive predictions, a non-negligible fraction are misclassified. Bayesian pooling corrects this local overconfidence and yields a more conservative estimate. This illustrates that very high confidence scores should not be equated with near certainty and that calibration-informed Bayesian adjustment exposes hidden residual risks.

Taken together, these four cases illustrate the complementarity of LPV and CPV. LPV captures local deviations from perfect calibration, while CPV provides a globally stabilized estimate. Their comparison highlights when local neighborhoods are misleading and when pooled evidence corrects over- or under-confidence.

4.3. Case study 3: PneumoniaMNIST

The PneumoniaMNIST test set consists of 624 pediatric chest X-ray images, sampled from a larger dataset of over 5000 images [25]. The task is framed as binary classification, distinguishing pneumonia from normal cases. For this dataset, we selected Auto-sklearn, which achieved an AUC of 0.937 and an accuracy of 0.853 on the original data [23]. Its performance on the binarized test dataset is reported in Tables 2 and 3, while its calibration plot is depicted in Fig. 8.

Case 1: Positive prediction with low confidence. This case highlights the tension between empirical frequency and calibration. The local neighborhood contained an overwhelming majority of positives (96 out of 100, $y_{\text{mean}} = 0.94$), which would naïvely suggest an almost perfect PPV. However, calibration was poor: the reference score (0.54) was far from the observed outcome rate, yielding a low local calibration index ($\text{ECI}_{\text{local}} = 0.27$). The LPV estimate was therefore moderate (0.83) with a very wide interval [0.14, 0.94]. The CPV mean was 0.79 ([0.75, 0.83]), tempering the misleading local evidence and producing a more conservative but reliable posterior. This illustrates how CIPV downweights bins that are empirically favorable but poorly calibrated (see Fig. 9).

Case 2: Negative prediction with low confidence. Here the classifier predicted a negative outcome with low confidence (close to the decision threshold). The local bin consisted entirely of negatives (100/100), which on its own would suggest perfect performance. Yet the calibration index was very low ($\text{ECI}_{\text{local}} = 0.13$), since the reference score (0.45) implied a much higher expected positive rate. Accordingly, the LPV estimate was 0.94 but with a very wide and uninformative interval [0.09, 1.00]. The CPV mean dropped to 0.77 ([0.71, 0.83]), showing how the framework avoids the illusion of certainty when calibration is poor, even in perfectly homogeneous bins. This example demonstrates one of CIPV's main strengths: integrating calibration quality with Bayesian pooling to protect against spurious local patterns and unjustified overconfidence.

4.4. Overall assessment

Across all benchmarks, the calibration-informed framework systematically moderated local empirical estimates according to two key determinants: the quality of local calibration (as measured by ECI) and the influence of global priors derived from the overall class distribution. When both calibration and empirical support were strong, the posterior estimates of the CPV closely reflected the local evidence. Conversely, when calibration was poor or when empirical frequencies were derived from narrow or unbalanced bins, the posterior shifted the CPV toward more conservative values. This mechanism ensures that the framework remains epistemically cautious while still responsive to reliable local information.

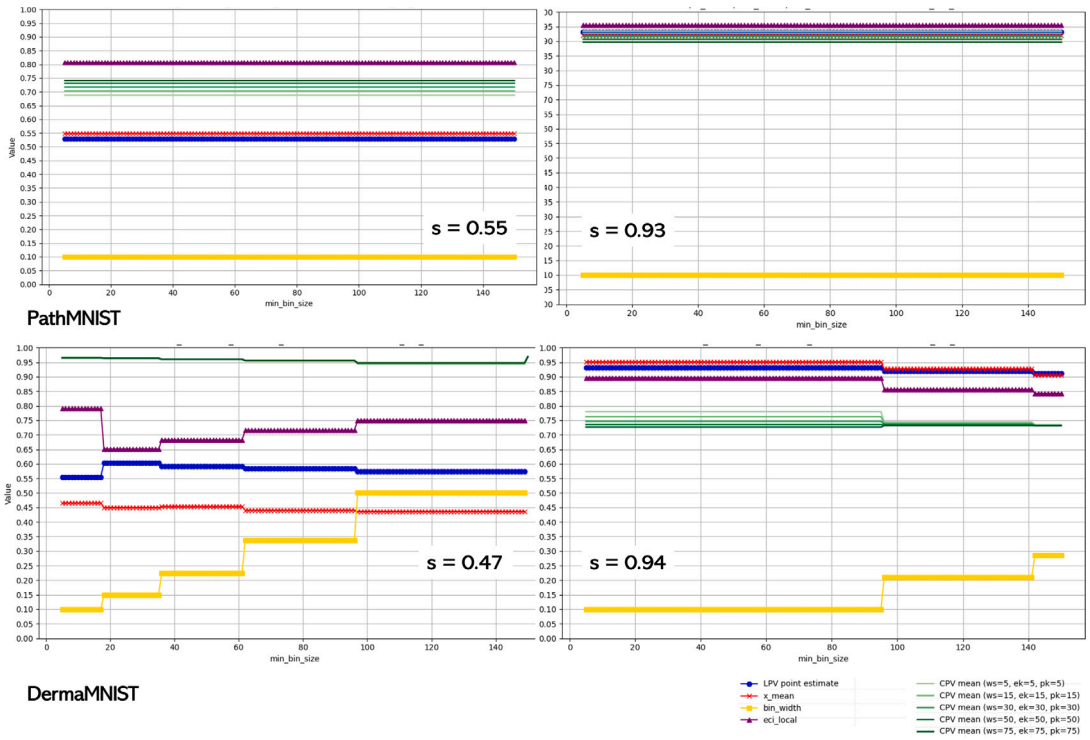


Fig. 7. Trends of calibration-related quantities as a function of the parameter *min_bin_size* in the PathMNIST dataset (top panels) and the DermaMNIST dataset (bottom panels), for two representative model confidence values (s, indicated within each panel). The curves display the LPV point estimate (blue), the average confidence within the bin (*x_mean*, red), the bin width (*bin_width*, yellow), the local calibration index (*eci_local*, purple), and the CPV mean computed under five distinct configurations of the hyperparameters (*weight_scale*, *evidence_kappa*, *pool_kappa*), shown in green with varying intensities. All variables are represented on the y-axis within the interval [0,1]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

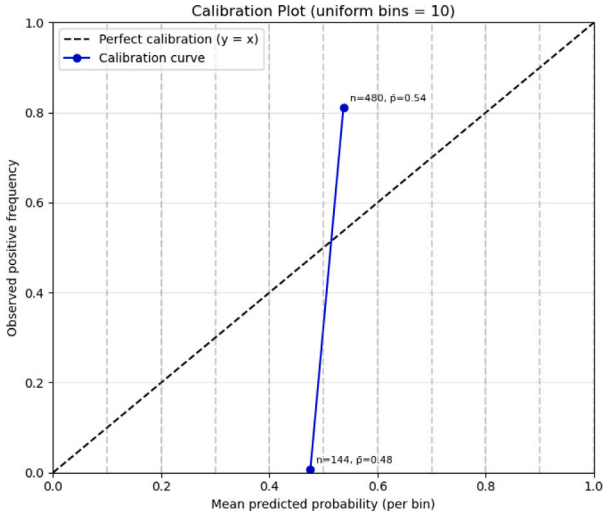


Fig. 8. Reliability diagram of the model evaluated on the pneumoniaMNIST benchmark with 10 uniform probability bins. For each bin, the annotation indicates the number of samples falling into that bin (n) and the average predicted probability within the bin ($\bar{p}$). The diagonal dashed line represents perfect calibration ($y = x$), while the blue curve shows the observed positive frequency as a function of the mean predicted probability. Vertical gray dashed lines denote bin boundaries.

Sensitivity to contextual parameters. All results above were obtained under a common set of modeling parameters that regulate how local and pooled information are combined. These parameters specify whether the effective weight derives from local evidence strength or from uncertainty reduction (cf. *weight mapping*); determine the overall contribution of the Bayesian update (cf. *weight scale*); and set the degree of shrinkage applied when local data are scarce (see *evidence kappa* and *pool kappa*). Taken together, these parameters define the balance between local calibration sensitivity and global stability, ensuring that both estimates remain interpretable and robust across different regions of the score spectrum.

In a limited number of cases — most notably, highly confident predictions with near-perfect local calibration — the posterior sometimes yielded estimates that appeared excessively conservative. Such outcomes highlight the potential need for fine-tuning meta-parameters (e.g., λ, κ) or extending the trust function ρ to account explicitly for variance.

This sensitivity to initial parameters is in itself desirable, as it allows the analyst to tailor the balance between local responsiveness and global robustness. However, we observed that the actual values of LPV and CPV are relatively stable across different parameter settings. Among the various parameters, the minimum bin size exerts the strongest influence, particularly when it is fixed rather than determined adaptively through the kNN method.

Fig. 7 illustrates how the main parameters, including LPV and CPV, change only slightly as bin size increases, even in the two extreme cases analyzed (PathMNIST and DermaMNIST benchmarks). Based on these observations, we conjecture that: (i) even small bins can yield meaningful LPV and CPV values; (ii) the smaller the parameters, the closer CPV tends to approach LPV; and (iii) regardless of their initial settings, the final estimates vary only within a narrow margin. Despite these nuances, the framework demonstrated strong internal coherence and conceptual defensibility across heterogeneous clinical classification settings.

5. Interactive tool for implementing the metric framework

To facilitate the practical adoption of the proposed metric framework in ML-based, AI-driven decision support systems, we developed a

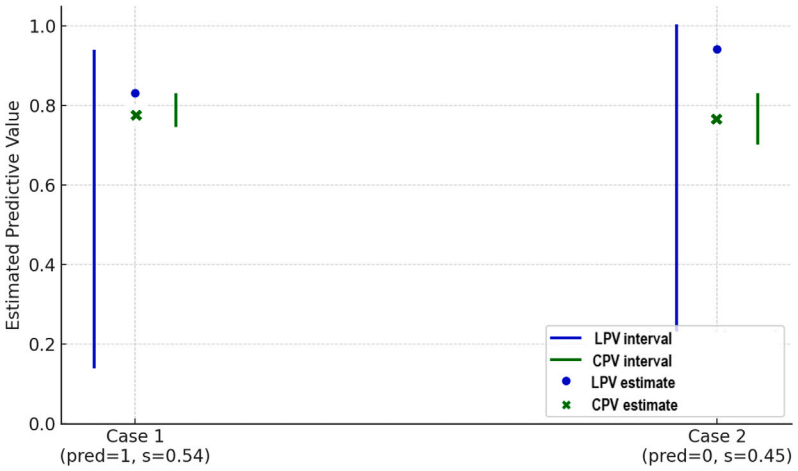


Fig. 9. Comparison of calibration-informed predictive value (CPV) and local predictive value (LPV) estimates for two representative cases for the pneumaMNIST benchmark. Blue markers indicate LPV point estimates with their corresponding uncertainty intervals (blue vertical lines), while green markers denote CPV point estimates with associated intervals (green vertical lines). For each case, the predicted label (*pred*) and the model's raw confidence score (s) are reported on the x-axis. The y-axis shows the estimated predictive value, highlighting differences in calibration-aware versus local empirical estimation of predictive reliability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

small interactive application that computes all the quantities defined in Section 2: this application is freely available at the following address: https://www.entechne.com/CIPR/[1]

The input page of the application allows the user to load or paste the evaluation data for a probabilistic classifier. The interface expects a test set consisting of a vector of predicted confidence scores, interpreted as estimated probabilities for the positive class, and the corresponding binary ground-truth labels (`y_prob` and `y_true`, respectively). Data can be imported from a comma-separated values (CSV) file or manually pasted into the text areas. In the lower panel, the user specifies the characteristics of the new case under consideration: the model prediction (0 or 1), the associated confidence score s^*, the decision threshold used in practice (e.g. 0.5), and the prevalence P of the target condition. Prevalence can either be estimated from the supplied test set or entered manually, to reflect an external epidemiological estimate. A further control lets the user select the neighborhood scope around s^*, mapping the abstract neighborhood definitions introduced in our framework (e.g. score-based or margin-based neighborhoods) to a concrete choice in the user interface.

Once these inputs are provided, the application computes the Calibration-Informed Predictive Value (CIPV) for the new case, together with the intermediate quantities defined in the framework. The first results view (Fig. 10(a)) reports the *Local Predictive Value* (LPV), i.e. the estimated probability that the model prediction is correct for cases with a similar confidence score and the same predicted label, together with a frequentist confidence interval obtained through the local empirical calibration procedure. It also reports the *Credible Predictive Value* (CPV), which combines the empirical calibration information with a Bayesian prior and is summarized by a posterior mean and a 95% credible interval. For both LPV and CPV, the interface includes short textual explanations that restate the interpretation of each quantity in non-technical terms, to support communication with clinicians and other stakeholders.

In addition, Fig. 10(a) illustrates a one-dimensional probability plot that juxtaposes four key elements on the same scale: (i) the baseline prevalence P; (ii) the CPV interval and its mean; (iii) the LPV interval and its mean; and (iv) the decision threshold used to dichotomize the probability into a positive or negative recommendation. This visual summary is intended to convey at a glance how much the model output, suitably calibrated and locally contextualized, changes the probability of the condition for the specific case at hand, compared with the population baseline.

A second results view focuses on calibration (Fig. 10(b)). Here the application displays a standard reliability diagram, plotting empirical outcome frequencies against predicted probabilities for the test set, together with the diagonal line of perfect calibration. The curve showing empirical calibration is augmented with a marker highlighting the local neighborhood of the new case: the horizontal position of the marker corresponds to the case's confidence score s^*, while its vertical position corresponds to the observed frequency of positive outcomes in the selected neighborhood. The shaded vertical band indicates the range of scores used to compute the local empirical calibration index (ECI), whose numerical value and contextual statistics (e.g. neighborhood size and event rate) are reported in the textual summary. The lower part of the view provides additional explanations of the concept of post-probability and its relation to the LPV, CPV, prevalence and decision threshold, again using non-technical language intended for end users.

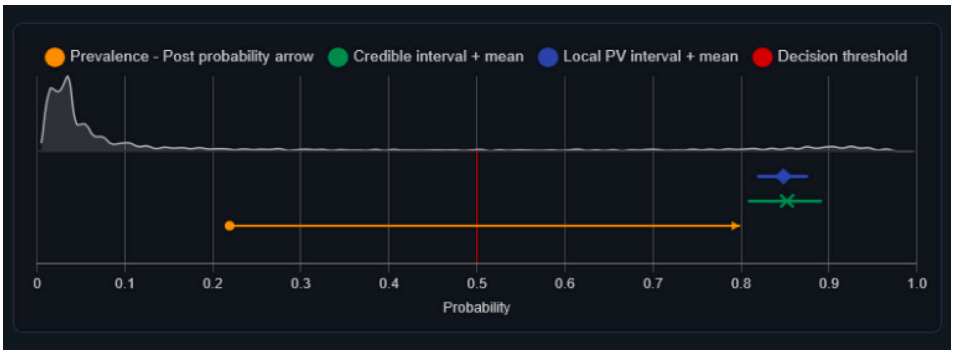
Overall, the application operationalises the metric framework in a way that is directly usable by practitioners, while keeping the framework itself independent of any specific software platform. Researchers and developers are free to embed the same computations in their own pipelines (e.g. during model evaluation or at prediction time in a clinical interface), and the tool presented in Fig. 10a–b should be regarded as an optional, reference implementation that illustrates how the proposed metrics can be presented and explained in an interactive setting.

6. Discussion and limitations

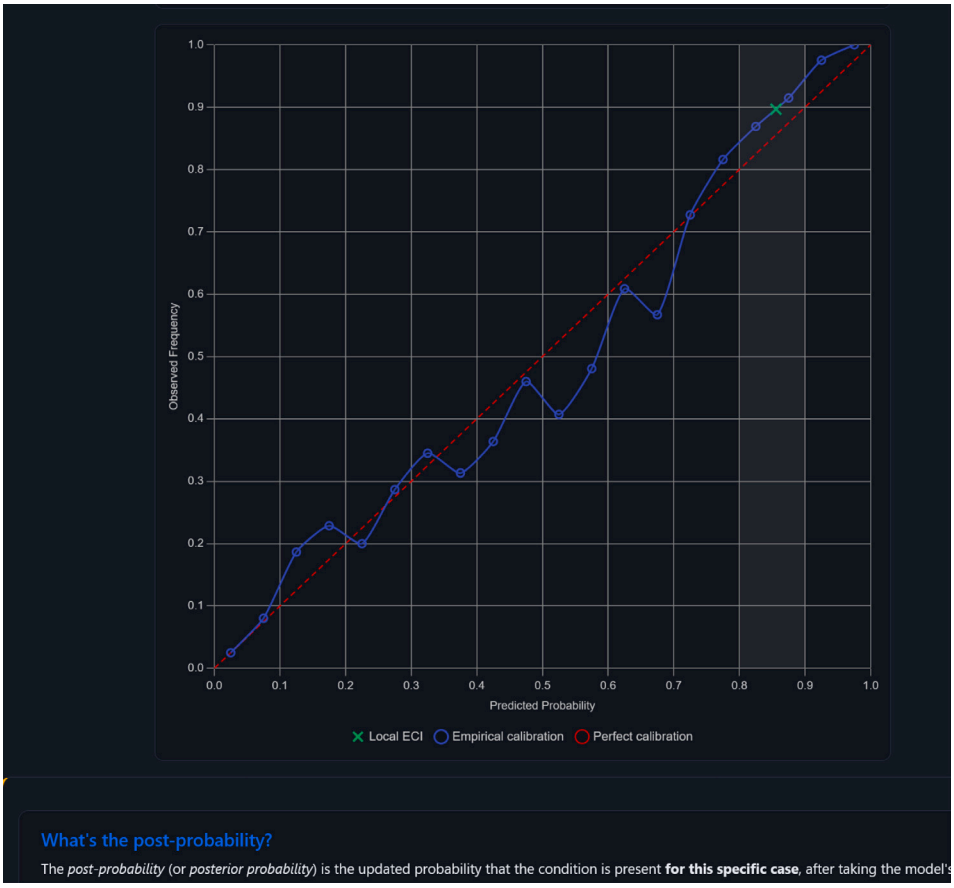
In this section, we discuss the implications and limitations of the proposed calibration-informed framework for instance-level reliability estimation in medical AI. We first summarize the main findings from the case studies, then relate our approach to existing work on calibration and uncertainty quantification, and finally outline its limitations and possible directions for future research.

Our metric framework addresses a critical gap in medical AI systems by providing a locally adaptive, calibration-sensitive measure of the reliability of individual predictions. Building on both local calibration

[1] Readers should notice that the framework itself is implementation-agnostic: any team can re-implement the same computations within its own software stack or clinical workflow. The application described here is therefore offered only as a reference implementation, *as is*, to make experimentation and dissemination easier, and it is not a prerequisite for using the framework.



(a) Graphical representation of the posterior probability metrics, LPV and CPV values and intervals.



(b) Reliability diagram illustrating observed frequency versus predicted probability.

Fig. 10. Data visualizations produced by the online tool: (a) detailed predictive metrics including uncertainty quantification, and (b) assessment of estimate reliability and calibration.

via the Estimated Calibration Index (ECI) [9] and global predictive distributions used as priors, the two proposed metrics refine the estimation of correctness probability for single model outputs. This design aligns with reported requirements from clinicians, who have emphasized the need for instance-specific reliability information.

The empirical results presented in this paper are based on three clinical imaging benchmarks, which we use as illustrative case studies rather than as a comprehensive validation across all possible applications or deployment settings.

In the context of clinical decision support, prior literature has established that the raw confidence scores of contemporary machine learning models are often miscalibrated [10,26] and that global metrics fail to convey the uncertainty of individual predictions. The importance of calibration is paramount, as it directly impacts the quality of clinical decisions derived from risk models [27].

As we will see in Section 6.1, various methods have been proposed to assess instance-specific trust, such as by analyzing a model's internal representations [7]; however, our approach differs from these methods in three main respects: We build upon recent work that highlights the importance of assessing calibration in a *local* neighborhood around a prediction [21], but our metrics take a crucial step further: instead of only measuring or correcting local calibration, they leverage it to derive two complementary, actionable, instance-specific predictive values. Our methodology is grounded in the principle that accurate calibration is a prerequisite for trustworthy uncertainty estimates [9, 28], which we operationalize through a novel integration of local diagnostics with Bayesian probabilistic modeling.

The Local Predictive Value (LPV) quantifies empirical reliability by estimating the observed frequency of correct outcomes among similar cases in the local neighborhood of the current prediction. This operationalizes trustworthiness in a frequentist sense: it answers the question, "How often has a prediction like this been correct?" and reflects the system's historical reliability in that region of the score space.

By contrast, the Credible Predictive Value (CPV) represents a calibrated belief in the correctness of a prediction, derived via Bayesian updating. It integrates prior information, typically global predictive values (PPV or NPV), with the evidence offered by the local outcome distribution, modulated by a credibility function that accounts for sample size and variance. The CPV thus expresses an epistemic judgment about the correctness of the current prediction, combining local and global perspectives. While the LPV embodies trust in the model's past behavior, the CPV reflects credence in the current prediction, given the available evidence. This distinction underscores the complementary roles of empirical reliability and probabilistic belief in evaluating prediction quality.

An important implication of this construction is that the local calibration index directly controls how much influence neighborhood evidence has on the posterior. When the local ECI is high, CPV strongly reflects local outcome frequencies. Conversely, when the ECI is low — indicating poor or noisy calibration — the posterior is shrunk toward the global prior predictive values. For practitioners, this means that CPV automatically tempers unreliable local evidence, ensuring that individual reliability estimates remain conservative rather than misleading.

Together, the Local and Credible Predictive Values provide complementary insights into the reliability of individual predictions. The LPV and its interval communicate the observed reliability of similar past predictions, which is especially informative when the local sample is sufficiently dense and well-calibrated. The CPV and its credible interval, by contrast, reflect a more holistic degree of belief in the prediction's correctness, tempering local empirical evidence with prior global knowledge and providing robustness in low-data or poorly calibrated regions. A clinician interpreting the LPV might assess whether the system has consistently succeeded or failed in comparable cases, while the CPV offers a belief-adjusted judgment of the present instance.

While local calibration may be noisy or discontinuous in practice, our framework explicitly mitigates this limitation: the Bayesian pooling in CPV smooths unstable local estimates by borrowing strength from global priors, thereby reducing the impact of non-representative or irregular bins

Importantly, LPV and CPV do not always align, and their divergence itself is informative [29]. A high LPV paired with a lower CPV, as seen in high-confidence positive predictions, may signal local overconfidence relative to global error rates. Conversely, a low LPV with a high CPV, often observed for confident negative predictions, indicates that global trends still support reliability despite noisy or unfavorable local evidence. Such discrepancies invite closer scrutiny, encouraging users to balance the descriptive view provided by LPV with the belief-adjusted judgment of CPV. By juxtaposing reliability and credibility, these dual estimates help mitigate both overreliance and unwarranted distrust, offering a nuanced view of the model's epistemic stance toward each decision [29].

The divergence between calibration-based and Bayesian estimates underscores a critical insight: even well-calibrated regions can yield overconfident raw predictions when global predictive values are substantially lower. This highlights the importance of including both local and global calibration considerations in reliability metrics. It also reveals potential discrepancies between a model's learned posterior and its empirical calibration behavior, pointing to avenues for improving calibration during model training.

Contributions and strengths. The proposed framework has four main strengths. First, it is model-agnostic and post-hoc: it can be applied to any probabilistic classifier, without access to internal parameters, multiple inference runs, or specific training pipelines. Second, it explicitly integrates local calibration information with global predictive priors, such as PPV, NPV, or disease prevalence. Third, it provides two complementary quantities, LPV and CPV, together with interpretable intervals that are closely aligned with familiar predictive values in medicine. Fourth, it is supported by an open Python implementation and a lightweight HTML interface, which make it straightforward to experiment with the metrics and to embed them in existing evaluation and decision-support workflows.

6.1. Comparison with other techniques for per-instance predictive uncertainty estimation

As we anticipated above, several established frameworks already aim to provide per-instance reliability estimates. Among the most prominent are *Bayesian approximation techniques*, *conformal prediction*, and *distance-based approaches* such as the Trust Score. While these methods share with our proposal the general goal of instance-specific predictive uncertainty estimation, they differ in key conceptual and practical aspects.

Bayesian approximations. Methods such as Monte Carlo Dropout [30] estimate epistemic uncertainty by keeping dropout layers active at inference and performing multiple stochastic forward passes. The resulting predictive distributions are aggregated to produce intervals or variances that reflect model uncertainty. While effective at capturing uncertainty in neural networks, these approaches are tightly coupled to specific architectures, require repeated inference for each test case, and yield estimates that are not explicitly connected to empirical calibration [10]. In practice, their computational demands and model dependence limit their usability in clinical deployment.

Conformal prediction. Conformal methods [31] construct prediction sets that contain the true label with a user-specified probability, providing strong, distribution-free guarantees under exchangeability assumptions. However, in the binary setting these sets often degenerate to either a single label or the full set $\{0, 1\}$, which severely restricts their informativeness for clinical decision-making [29]. Moreover, conformal prediction typically returns sets rather than calibrated probabilities, which reduces its interpretability in contexts where decision-makers require a single, actionable reliability estimate.

Distance-based methods. The Trust Score [7,17] represents a different line of work, measuring the plausibility of a prediction by comparing the distance of a test instance to training examples of the predicted class relative to alternative classes. These approaches are appealingly model-agnostic and useful for detecting out-of-distribution inputs, but they rely on feature-space distances that may be ill-defined or difficult to interpret in high-dimensional clinical data [32]. More importantly, they provide signals of instance plausibility rather than a calibration-informed estimate of prediction correctness.

By contrast, the Local Predictive Value (LPV) and Credible Predictive Value (CPV) proposed in this work differ in three crucial respects. First, they are *model-agnostic*: they can be applied to any classifier, regardless of its internal architecture. Second, they are *post-hoc and user-side*: they require only minimal information that vendors already disclose (true labels and predicted probabilities on a test set), without access to model parameters, multiple inference runs, or training pipelines. Third, they are *explicitly calibration-informed*: LPV integrates local empirical calibration into a reliability estimate, while CPV further incorporates Bayesian pooling against global predictive values (PPV and NPV), producing tempered and interpretable posterior estimates.

These characteristics confer several advantages. (i) *Practical feasibility*: LPV and CPV can be computed directly from data typically available

in model cards or evaluation reports, facilitating adoption in regulated settings. (ii) *Interpretability*: unlike many explainability (XAI) methods that have been criticized as insufficient for clinical practice [33,34], LPV and CPV provide intuitive, probabilistic quantities that directly answer the pragmatic question: "How much can I trust *this specific* prediction?" (iii) *Robustness*: by combining local calibration with global priors, CPV guards against misleading local evidence and produces conservative, clinically actionable reliability intervals. By quantifying calibration uncertainty in a transparent and decision-relevant way, the proposed framework supports trustworthy decision support, where clear communication of uncertainty informs clinical judgment and shared decision-making.

In sum, while Bayesian approximations, conformal prediction, and distance-based methods offer important contributions to the field of predictive uncertainty estimation, LPV and CPV represent a distinct and complementary family of methods. They are especially well aligned with the epistemic, practical, and regulatory requirements of clinical decision support, where actionable, case-specific, and calibration-informed reliability estimates are paramount.

6.2. Limitations and future works

While we made the point that the Local Predictive Value (LPV) and Credible Predictive Value (CPV) offer promising tools for per-instance predictive uncertainty estimation, several limitations and open directions deserve attention. We outline them in what follows, and then we describe corresponding avenues for future investigation for the most important open limitations.

1. the dependence of local estimates on sufficient neighborhood support and the need to better characterize the behavior of the framework under extreme sparsity;
2. open questions about robustness under distribution shift and adversarial manipulations.
3. the current focus on binary classification, with multiclass and other settings left for future work;
4. the absence, so far, of systematic user studies with clinicians interacting with LPV and CPV in realistic decision-making scenarios;
5. the use of retrospective, controlled benchmark datasets from the MedMNIST collection rather than prospective, real-time clinical deployments;

First, we acknowledge that further investigation is needed into the behavior of LPV and CPV under adverse conditions. These include scenarios of *model drift*, where changes in patient populations or data distributions may undermine the validity of previously estimated predictive values [15], as well as adversarial manipulations that could distort confidence scores. Understanding the robustness of calibration-informed reliability estimates in such settings is essential for their safe adoption. However, these concerns relate more to the frequent updating of the validation/test datasets — and thus to the currentness and quality of the priors used in our estimators — than to the estimators themselves. Accordingly, this limitation points to a direction for future work. Moreover, because the framework is local by construction, LPV and CPV inherently depend on having sufficient neighborhood support around a given score; in extremely sparse regions of the score space, the resulting estimates may become unstable despite the pooling mechanisms we employ, and warrant further investigation.

Second, integration with other approaches to model transparency remains an open research avenue. While LPV and CPV directly address the pragmatic question of prediction reliability, they could be complemented by established explainability (XAI) methods to provide richer, multi-faceted accounts of model behavior. In this respect, we suggest that the integration of distance-based techniques constitutes a promising research direction. However, it should be noted that, in binary classification, confidence scores near the commonly used threshold of 0.5 already serve as an internal — albeit indirect — indication that the training data may not be strongly representative of the given instance, effectively reflecting the instance's distance from the training distribution. Exploring such combinations may yield tools that are both interpretable and trustworthy in high-stakes decision support.

Third, methodological extensions are needed. At present, LPV and CPV are defined for binary classification. Extending the framework to multiclass settings and regression tasks would broaden its applicability to a wider range of medical AI systems. Similarly, the design of adaptive pooling strategies — balancing local empirical frequencies with global priors in a task-sensitive manner — could further enhance reliability estimates. That said, binary outcomes remain informative in multiclass classification and regression tasks, where clinically plausible thresholding is common practice, and both types of problems can be reduced to a series of binary decisions, when appropriate. Moreover, we have already discussed the limitations of distance-based measures (see Section 6.1), particularly with regard to human plausibility and interpretability [32]. For these reasons, we did not consider extending the framework to multiclass and regression tasks, or integrating distance metrics, to be urgent at this stage, but rather a direction for future work.

Lastly, our evaluation has been limited to retrospective, controlled benchmark datasets drawn from the MedMNIST collection. Future research should investigate the deployment of LPV and CPV in real-world clinical workflows, assessing how case-specific reliability estimates influence physicians' trust in the system, diagnostic accuracy, decision quality, appropriate reliance [35], and ultimately, patient outcomes. Such empirical studies are essential to determine whether the proposed metrics enhance interpretability, user understanding, and confidence in decisions. To our knowledge, however, no existing performance metric has been evaluated through such an empirical user study. While this gap also applies to our CIPVs, it has not deterred us from proposing this methodological contribution.

In summary, while LPV and CPV provide a conceptually simple yet powerful framework for instance-level reliability estimation, future work should address their validation in clinical environments, their robustness to distributional and adversarial shifts, their integration with complementary interpretability techniques, and their extension beyond binary classification. Addressing these challenges will be key to realizing the full potential of calibration-informed reliability measures in the design of safe and trustworthy AI systems for healthcare.

6.3. Recommendations for providers

Our two proposed metrics also imply a different interpretation of the general requirement for transparency in machine learning models and, more broadly, in AI-based decision support systems [36]. Article 13 of the EU AI Act, for instance, stipulates that deployers, i.e., those who operate such systems under their own authority, must be informed of the level of accuracy that can reasonably be expected from the system [37], and must implement oversight measures that facilitate the interpretation of outputs from high-risk AI systems (such as those used in healthcare). More specifically, Article 14(d) requires that a system provided to a deployer must allow them "to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system".

Compliance with these obligations is commonly understood as the provision of model documentation (e.g., model cards [38,39]) that includes a range of what could be considered *vanity measures* [13,40]: marginal classification and discrimination metrics derived from confusion matrices, typically reported on external validation datasets of uncertain representativeness for the deployer's own data.

Our approach requires providers to go further. It is not sufficient to expose only a global model description; models should also provide (i) the confidence scores associated with each new instance, and (ii) the

true labels (y_{true}) and corresponding confidence scores (y_{prob}) produced on a test or validation dataset. Such information need not compromise intellectual property, since test features may remain withheld and only target variables (or their numeric proxies) are needed. At the same time, it enables the application of the procedures described in this work, thereby yielding instance-specific predictive values.

This practice would give concrete meaning to the rationale of the regulation and to the broader requirement of transparency and interpretability [2,36,41,42]. The methods presented here thus also serve as an encouragement for providers to share, and for deployers and users to demand, the minimal data needed to support a more faithful and practical assessment of predictive reliance in AI-driven decision support systems.

6.4. Recommendations for deployers and practitioners

In clinical and high-stakes contexts where decisions must be made rapidly yet responsibly, we recommend the use of the *Credible Predictive Value* (CPV) posterior mean as the primary summary statistic of model reliability at the instance level. This value, typically expressed as a percentage (e.g., 0.805, or 81%), synthesizes both global predictive performance and local empirical support, offering a calibrated belief in the correctness of the prediction at hand. Its interpretability as a degree of credence makes it particularly suitable for communication in risk-sensitive settings.

To appropriately convey residual uncertainty around this estimate, we further advise reporting the associated *credible interval* (e.g., [0.765, 0.842]), which quantifies the plausible range of values the true predictive value might take, given the observed evidence. This allows decision-makers to assess not only the expected reliability of the prediction but also the level of confidence the system places in that assessment. For instance, a wide interval may suggest caution or the need for further review, while a narrow one may reinforce trust in the prediction's validity.

However, in contexts where understanding the *empirical* behavior of the model is also important — such as post-hoc auditing, model validation, or when assessing generalizability to new populations — the *Local Predictive Value* (LPV) and the size and cardinality of the local bin should be explicitly reported. The LPV reflects the historical correctness rate of similar predictions in the local neighborhood of the confidence space, thereby serving as a measure of the model's contextual trustworthiness. Reporting both LPV and CPV allows practitioners to disentangle situations in which the model is empirically reliable from those in which its prediction is credible due to broader prior performance but lacks sufficient local support.

In sum, we recommend a layered approach to reliability reporting: use the CPV posterior mean for decision support, its credible interval to communicate epistemic uncertainty, and the LPV and local sample characteristics to inform judgments about contextual trust. Such multi-faceted reporting can enhance transparency, mitigate automation bias, and promote safer and more transparent human–AI interaction in clinical workflows.

7. Conclusion

This work addresses a critical limitation in current machine learning evaluation practices by introducing a calibration-informed framework that bridges a methodological gap in predictive model evaluation. Using three publicly available clinical imaging benchmarks as illustrative case studies, we showed that the proposed metrics can be applied to real diagnostic classifiers and yield case-specific reliability estimates. We presented two novel metrics—the Local Predictive Value (LPV) and the Credible Predictive Value (CPV), designed to estimate the reliability of individual predictions in binary classifiers. Unlike conventional global metrics, these measures integrate the model's local calibration behavior with global predictive values through a combination of empirical calibration, global priors, and Bayesian inference, yielding correctness probabilities that are both statistically grounded and clinically interpretable.

By applying these metrics to real-world benchmark datasets and diagnostic binary classifiers, we showed that, in our case studies, they exhibit plausible and interpretable behavior and can generate locally adaptive reliability estimates by balancing information from empirical calibration (via the Estimated Calibration Index) and prior predictive distributions. The confidence-based (LPV) and Bayesian (CPV) variants provide complementary insights, showing how local evidence and global priors can be harmonized to enhance prediction trustworthiness.

These findings have implications for human-centered AI in healthcare. We hypothesize that the proposed calibration-informed scores can help clinicians better assess the reliability of individual predictions, thereby supporting more informed and context-sensitive decision-making; however, this potential benefit requires empirical confirmation through dedicated user studies in real clinical workflows. The results presented here makes the framework a promising addition to the methodological toolbox for building trustworthy AI systems.

A key limitation of the framework lies in its reliance on sufficient local data for robust calibration estimation. Nonetheless, the Bayesian approach offers a principled mechanism to mitigate this limitation by shrinking estimates toward global priors. Future work should examine the performance of LPV and CPV under distributional shift, adversarial conditions, and in multi-class classification scenarios.

In conclusion, Calibration-Informed Predictive Values, and specifically Local Predictive Values and the Credible Predictive Values—represent a significant methodological advancement toward the development of AI systems that are not only accurate but also reliably interpretable at the level of individual predictions, yielding reliable, interpretable, and context-sensitive measures of prediction correctness in healthcare and other high-stakes domains.

Declaration of competing interest

The author have no conflicts of interest to declare.

Acknowledgments

FC acknowledges funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

Appendix. Implementation and configuration details

The CIPR framework has been implemented as a self-contained Python script for post-hoc computation of Calibration-Informed Predictive Values (CIPV), which is publicly available together with an example command-line interface.[2] A corresponding client-side HTML/JavaScript version is also provided for privacy-preserving use in clinical settings.[3] In all cases, the framework operates on precomputed model outputs and does not require retraining or modification of the underlying predictive models.

From the standpoint of reproducibility, it is important to distinguish two separate aspects: (i) the *CIPV computation*, i.e. how LPV and CPV are derived from model scores and true labels; and (ii) the *training of the predictive models* used in our case studies. The present work focuses on (i): given a classifier that produces probabilistic predictions for a binary outcome, the script takes as input its scores and associated ground-truth

[2] Currently hosted at https://www.entechne.com/CIPR/CIPR-framework-script.zip.

[3] See https://www.entechne.com/CIPR/.

labels and returns per-instance LPV and CPV estimates with uncertainty intervals. The convolutional neural networks used for the case studies are standard models trained on publicly available datasets (PathMNIST, DermaMNIST, and PneumoniaMNIST from the MedMNIST collection, which is publicly available.), and can in principle be replaced by any other probabilistic classifier producing calibrated scores in $[0, 1]$.

The Python implementation, from which the JavaScript version is derived, comprises the following core components:

- `local_eci_from_reference_confidence_L1`: estimates the local Estimated Calibration Index ($\mathrm{ECI}_{\mathrm{local}}$) for a reference score using k-nearest neighbors or adaptive binning in confidence space, based on vertical deviation from perfect calibration.
- `_beta_credible_interval`: computes Bayesian credible intervals for Beta distributions (if SciPy is available).
- `compute_cipv`: calculates both deterministic (LPV, from local calibration) and Bayesian (CPV, from a posterior over predictive values) estimates for a single instance.
- `cipv_from_local_eci`: a wrapper that integrates local ECI estimation with CIPV computation, derives local and global confusion counts, and returns a structured result object.
- `format_cipv_report`: generates a human-readable summary of results, including contextual parameters for both LPV and CPV.

CIPV computation pipeline

Given a trained probabilistic classifier and a validation dataset with its outputs, the computation of LPV and CPV for a new instance follows the steps implemented in the Python script:

1. **Prepare the validation data.** A CSV file is provided containing, for each validation case, the model-estimated probability of the positive class $P(Y{=}1 \mid X)$, the ground-truth label, and optionally the predicted label. If the predicted label is not available, it is reconstructed by thresholding the probability at a configurable threshold (default: 0.5).
2. **Specify the target instance.** For the new case of interest, the user supplies the model's predicted label (1 for the PPV branch, 0 for the NPV branch) and the corresponding confidence score in $[0, 1]$.
3. **Select the local neighborhood in score space.** Using the validation data, a local neighborhood around the target confidence is built in score space. The script supports:
   - a legacy *margin* mode, which grows a symmetric interval around the reference confidence until a minimum number of neighbors or a maximum margin is reached;
   - a *k-NN* mode (default), which selects the k closest validation scores to the reference confidence and optionally applies a kernel (uniform, triangular, or Gaussian) to weight neighbors.

   An additional scope parameter allows restricting the neighborhood to cases sharing the same predicted label as the target instance.
4. **Estimate local calibration (local ECI).** Within the selected neighborhood, the script computes the mean confidence and the empirical outcome rate, and derives a local Estimated Calibration Index $\mathrm{ECI}_{\mathrm{local}}$ based on their vertical deviation from the diagonal $y = x$ in the reliability plane. The function returns both $\mathrm{ECI}_{\mathrm{local}}$ and a detailed context object containing neighborhood size, bin bounds, local means, and other diagnostics.
5. **Compute local and global confusion counts.** From the validation set, the script constructs the global confusion matrix $(\mathrm{TP}, \mathrm{FP}, \mathrm{TN}, \mathrm{FN})$. The same counts are also computed within the local neighborhood, yielding bin-level counts $(\mathrm{TP}_{\mathrm{bin}}, \mathrm{FP}_{\mathrm{bin}}, \mathrm{TN}_{\mathrm{bin}}, \mathrm{FN}_{\mathrm{bin}})$ and a local bin size.
6. **Derive the LPV estimate and its interval.** Using $\mathrm{ECI}_{\mathrm{local}}$, the local event rate, and the target confidence, the script computes a deterministic LPV quantity (ECI-only predictive value) and a corresponding interval around the score that reflects local calibration uncertainty. This step yields a frequentist, calibration-informed predictive value for the specific instance.
7. **Derive the CPV estimate and its credible interval.** When global confusion counts are available, the script performs a Bayesian update on a Beta prior over predictive values, optionally pooling global and local information. The mapping from local calibration to statistical weight is controlled by user-selectable parameters (weight mapping, scaling, and shrinkage towards a pooled prior). The result is a posterior mean CPV and a $(1 - \alpha)$ credible interval.
8. **Produce a human-readable report.** Finally, the script can generate a textual report summarizing LPV and CPV, their intervals, and key elements of the local calibration context, which can be presented to end users or stored for audit purposes.

Input data and context parameters

To compute LPV and CPV for a new instance, two types of data are required. First, a validation set containing the model's predicted probabilities for the positive class (y_{prob}) and the corresponding ground-truth labels (y_{true}). This dataset is used to estimate local calibration and to derive prior information for Bayesian pooling. Second, for the instance under evaluation, the model's predicted probability (y_{prob}) and, optionally, its binary predicted label (y_{pred}). If not provided, y_{pred} is reconstructed by thresholding y_{prob} at 0.5. The script therefore assumes that `-conf-col` corresponds to the predicted probability of the positive class, $P(Y{=}1|X)$, while `-pred-col` is optional.

LPV context. The local context includes:

- the number of neighboring cases (k),
- the kernel and bandwidth used to weight neighbors (e.g. triangular kernel with adaptive bandwidth),
- the local $\mathrm{ECI}_{\mathrm{local}}$, quantifying deviation from perfect calibration,
- the bin interval and effective margin, defining the neighborhood around the reference confidence,
- the empirical outcome rate and mean confidence in the neighborhood, revealing local model bias.

CPV context. The Bayesian posterior additionally depends on:

- *weight mapping* (e.g. `evidence`), which determines whether local reliability is mapped from evidence strength or uncertainty reduction;
- *weight scale mode* (e.g. `const`) and the scaling factor λ, which regulate the strength of the Bayesian update;
- *evidence kappa*, stabilizing the mapping from local evidence to statistical weight;
- *pool kappa*, which controls the shrinkage towards the pooled prior;
- *use pooled prior*, enabling the model to borrow global information when local data are weak;
- *credible level* (default: 0.95), defining the width of the Bayesian credibility interval.

These hyperparameters were chosen to provide a balanced trade-off: sufficient pooling to prevent instability in regions with sparse or poorly calibrated local evidence, while still allowing local calibration signals to influence the posterior meaningfully. Lower values of λ or κ make the posterior more responsive to local fluctuations, whereas higher values enforce stronger shrinkage towards global priors. Unless otherwise stated in the text, all case studies reported in Section 4 use the default configuration summarized in Table A.4, with k-nearest neighbor

Table A.4
Input parameters for LPV and CPV estimation, with default values and interpretation.

Parameter	Default	Meaning/Role
LPV context		
k (neighbors)	50	Number of nearest neighbors considered in local calibration estimation.
Kernel	triangular	Weighting function applied to neighbors, emphasizing proximity in confidence space.
Bandwidth h	adaptive	Effective margin around the reference confidence; defines neighborhood size.
$\text{ECI}_{\text{local}}$	–	Deviation from perfect calibration in the local neighborhood.
Bin interval	data-driven	Range of confidence values forming the neighborhood.
Empirical outcome rate	–	Observed frequency of positive outcomes in the neighborhood.
Mean confidence	–	Average predicted probability in the neighborhood.
CPV context		
Weight mapping	evidence	Determines whether local reliability derives from evidence strength or uncertainty reduction.
Weight scale mode	const	Regulates scaling of Bayesian update (`const` vs. `bin_fraction`).
Weight scale λ	30.0	Magnitude of weight applied to local evidence in the Bayesian update.
Evidence kappa	30.0	Stabilizer for the mapping from local evidence to statistical weight.
Pool kappa	30.0	Shrinkage parameter controlling the strength of the pooled prior.
Use pooled prior	True	Enables borrowing strength from the global distribution when local evidence is weak.
Credible level	0.95	Bayesian credible interval level.

neighborhoods in score space, a triangular kernel, an evidence-based weight mapping in the Bayesian update, a constant weight scale, pooled priors enabled, and a credible level of 0.95.

Input Checklist

To compute the LPV and CPV in practice, the following inputs are needed:

- **From a hold-out or external validation set:**
  - Model's predicted labels.
  - Associated confidence scores (probabilities) for the positive class.
  - Ground-truth outcomes (true labels).

  Purpose: estimate local calibration (ECI) and compute global or pooled confusion counts for priors. *Note:* these data should be provided by the model developer, e.g. in the model card or technical documentation. In our case studies, these validation sets are derived from the PathMNIST, DermaMNIST, and PneumoniaMNIST subsets of the MedMNIST collection.
- **From the new case of interest:**
  - Model's predicted label (positive/negative), optional.
  - Associated probability or confidence score, mandatory.

  Purpose: generate the individualized predictive values for that prediction.

With these elements, all predictive values defined above can be produced: the local predictive value estimate and its reliability interval, and the credible predictive value with its credible interval.

References

[1] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, Weicken E, Asselbergs FW, Prior F, Collins GS, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. Bmj 2025;388.

[2] Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell 2022;5:879603.

[3] Busch F, Kather JN, Johner C, Moser M, Truhn D, Adams LC, Bressem KK. Navigating the European union artificial intelligence act for healthcare. Npj Digit Med 2024;7(1):210.

[4] Yang CC. Explainable artificial intelligence for predictive modeling in healthcare. J Heal Informatics Res 2022;6(2):228–39.

[5] Sivaraman V, Bukowski LA, Levin J, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In: Proceedings of the 2023 CHI conference on human factors in computing systems. 2023, p. 1–18.

[6] Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. PLoS One 2020;15(2):e0229132.

[7] Jiang H, Kim B, Guan MY, Frosst N. To trust or not to trust a classifier. In: Advances in neural information processing systems, vol. 31, 2018.

[8] Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, Tran D, Lucic M. Revisiting the calibration of modern neural networks. Adv Neural Inf Process Syst 2021;34:15682–94.

[9] Famiglini L, Campagner A, Cabitza F, et al. Towards a rigorous calibration assessment framework: advancements in metrics, methods, and use. Frontiers Artificial Intelligence Appl 2023;372:645–52.

[10] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th international conference on machine learning, vol. 70, JMLR. org; 2017, p. 1321–30.

[11] Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev 2023;56(Suppl 1):1513–89.

[12] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion 2021;76:243–97.

[13] Cabitza F, Campagner A, Datteri E. To err is (only) human. Reflections on how to move from accuracy to trust for medical AI. In: Exploring innovation in a digital world: cultural and organizational challenges. Springer; 2021, p. 36–49.

[14] Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015;35(2):162–9.

[15] Chi S, Tian Y, Wang F, Zhou T, Jin S, Li J. A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. Artif Intell Med 2022;125:102256.

[16] Sternini F, Ravizza A, Cabitza F, et al. How accurate do you want it? Defining minimum required accuracy for medical artificial intelligence. Proc E-Health 2020.
[17] Cabitza F, Campagner A, Soares F, de Guadiana-Romualdo LG, Challa F, Sulejmani A, Seghezzi M, Carobene A. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed 2021;208:106288.
[18] Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, Cato KD. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: qualitative descriptive study. JMIR Hum Factors 2022;9(2):e33960.
[19] Pfohl SR, Foryciarz A, Shah NH. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. Sci Rep 2022;12(1):3254. http://dx.doi.org/10.1038/s41598-022-07167-7.
[20] Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. Jama 2018;320(21):2199–200.
[21] Luo R, Bhatnagar A, Bai Y, Zhao S, Wang H, Xiong C, Savarese S, Schmerling E, Pavone M. Local calibration: Metrics and recalibration. In: The 38th conference on uncertainty in artificial intelligence. 2022.
[22] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, Gaiser T, Marx A, Valous NA, Ferber D, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med 2019;16(1):e1002730.
[23] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Sci Data 2023;10(1):41.
[24] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5(1):1–9.
[25] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19(1):221–48.
[26] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the achilles heel of predictive analytics. BMC Medicine 2019;17(1):230. http://dx.doi.org/10.1186/s12916-019-1466-7.
[27] Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015;35(2):162–9.
[28] Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. In: International conference on machine learning. PMLR; 2018, p. 2796–804.
[29] Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis.. Artif Intell Med 2024;150:102830.
[30] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. PMLR; 2016, p. 1050–9.
[31] Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. Found Trends Mach Learn 2023;16(4):494–591.
[32] Cabitza F, Famiglini L, Campagner A, Sconfienza LM, Fusco S, Caccavella V, Gallazzi E. Dissimilar similarities: Comparing human and statistical similarity evaluation in medical AI. In: International conference on modeling decisions for artificial intelligence. Springer; 2024, p. 187–98.
[33] Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency. 2019, p. 279–88.
[34] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745–50.
[35] Cabitza F, Campagner A, Fregosi C, Cameli M, Gallazzi E, Sconfienza LM, Tontini GE. Five degrees of separation: Investigating the unexpected potential of displaced human-AI collaboration protocols for apter AI support. In: Proceedings of the 28th ACM SIGCHI conference on computer-supported cooperative work and social computing. CSCW'25, Association for Computing Machinery; 2025, http://dx.doi.org/10.1145/3757601.
[36] Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, Holmes JH. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423.
[37] Regulation (EU) 2024/1689 of the European parliament and of the council (artificial intelligence act). Off J Eur Union 2024;L 2024/1689. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.
[38] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T. Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. ACM; 2019, p. 220–9. http://dx.doi.org/10.1145/3287560.3287596, URL https://dl.acm.org/doi/10.1145/3287560.3287596.
[39] Pushkarna M, Zaldivar A, Kjartansson O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 2022, p. 1776–826.
[40] Selbst AD, Barocas S. The intuitive appeal of explainable machines. Fordham L. Rev 2018;87:1085.
[41] Högberg C, Larsson S. AI and patients' rights: Transparency and information flows as situated principles in public health care. In: De lege–yearbook uppsala faculty of law 2021: law, AI & digitalization. Iustus förlag; 2022, p. 401–29.
[42] Gilbert S. The EU passes the AI act and its implications for digital medicine are unclear. NPJ Digit Med 2024;7(1):135.