Computers and Electronics in Agriculture 237 (2025) 110559

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty quantification

Mohamed Farag [a,*], Ahmed Emam [a], Johannes Leonhardt [a], Ribana Roscher [b,a,**]

[a] Remote Sensing Group, Institute of Geodesy and Geoinformation, University of Bonn, Germany
[b] Data Science for Crop Systems, Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Germany

ARTICLE INFO

Keywords:
Digital agriculture
Crop production
Deep learning
Uncertainty quantification
Conformal prediction

ABSTRACT

Assessing the confidence of machine learning models is increasingly important for enhancing decision support in digital agriculture. Addressing this challenge involves identifying and classifying sources of uncertainty and applying a suitable quantification method. To fully realize the potential of uncertainty quantification in digital agriculture, it is necessary to explore and evaluate different techniques, demonstrating their effectiveness in diverse scenarios. In this paper, we focus on the conformal prediction (CP) framework, specifically inductive conformal prediction (ICP), as a non-parametric tool for uncertainty quantification. Conformal prediction has gained a foothold for its multiple advantages, including being model-agnostic – requiring no retraining or changes to model architecture – and computationally efficient. Moreover, CP operates under a distribution-free framework, making no assumptions about the data, and provides calibrated uncertainty estimates where empirical errors match pre-defined theoretical levels. These properties make it particularly well-suited for agricultural tasks, where datasets often exhibit variability across regions and lack reliable ground truth labels. Through ablation studies, we comprehensively analyze ICP's performance as a post-hoc approach for ResNet-18 and ViT-B/16, demonstrating that both achieve the required pre-defined error levels. We compare ICP performance against other methods in diverse agricultural tasks, including out-of-distribution detection and covariate shift. Our experiments highlight CP's unique benefits, such as validity, efficiency, and low computational overhead, making it a promising approach for developing more reliable agricultural machine learning systems.

1. Introduction

Despite the United Nations (UN) introducing the *Zero hunger* goal in 2012 to eradicate hunger worldwide by 2030, approximately 900 million people still face severe food insecurity (FAO et al., 2022). This shortfall underscores the urgent need to improve our agricultural strategies to ensure long-term food security for all. The rise of digital agriculture as a solution has gained popularity by its embrace of data-driven models, which are built upon the observations from sensors and strategies used by Machine learning (ML). Such models have the potential to not only enhance existing agricultural management practices but also to shed light on viable new strategies for future management.

Particularly, Deep Learning (DL) has proven to be exceptionally effective in processing large datasets collected via various sensors, identifying significant patterns and insights in Pound et al. (2017). This capability has paved the way for a diverse array of applications, including the classification of crops using multiple models (Teixeira et al., 2023), the modeling and forecasting of plant growth stages (Drees et al., 2022, 2024; Quan et al., 2019), detection of plant diseases (Ferentinos, 2018; Neupane et al., 2019), and yield estimation (Zabawa et al., 2020).

In digital agriculture, the growing reliance on ML models makes it essential to evaluate prediction quality. Uncertainty quantification offers valuable insights into model outputs, boosting confidence and signaling when human intervention is needed. In other words, instead of just generating *point-wise estimates*, uncertainty tools as shown in Fig. 1 allow to investigate issues inherited by data, model, or both by producing a *probabilistic* prediction.

ML models, in which labeled data is used to train a model to make predictions on unseen data, is considered learning by induction. As such, models approximate the real world, creating a source of uncertainty, denoted as *Epistemic* (or model uncertainty). Additionally, noisy and imprecise data serve as another source of uncertainty, denoted as

* Corresponding author.
** Correspondence to: Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, 52428 Jülich.
E-mail addresses: mibrahi2@uni-bonn.de (M. Farag), ribana.roscher@uni-bonn.de (R. Roscher).

https://doi.org/10.1016/j.compag.2025.110559
Received 15 October 2024; Received in revised form 28 February 2025; Accepted 13 May 2025
Available online 6 June 2025

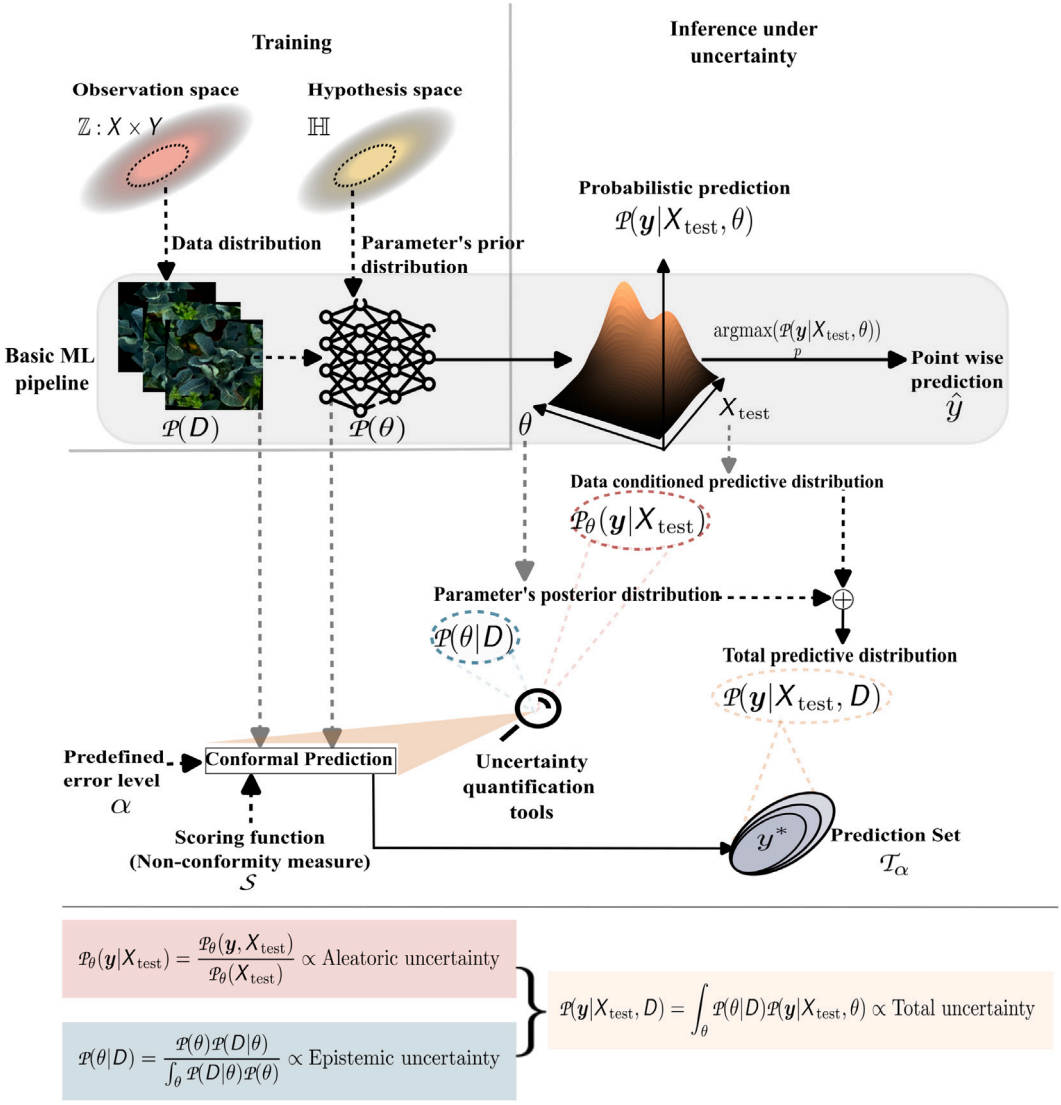


Fig. 1. Uncertainty quantification in Machine learning. During training, we acquire finite samples D from the original observational space $\mathbb{Z}$. We select our hypothesis space $\mathbb{H}$, and based on the prior parameters $\mathcal{P}(\theta)$ and data distribution $\mathcal{P}(D)$, we obtain our model $\mathcal{F}$. *Basic machine learning pipeline* focuses on acquiring the label with the highest probability $\hat{y}$, which is a point-wise prediction (*simple prediction*). To ensure reliability, we switch to *Probabilistic* inference to quantify uncertainties from the model and new data points, represented by the parameters' posterior distribution $\mathcal{P}(\theta|D)$ and conditional $\mathcal{P}_\theta(y|X_{\text{test}})$ distributions, respectively. Uncertainty quantification techniques provide tools to measure data, model, and total predictive uncertainties. Conformal prediction as a post-hoc approach uses data, scoring function $\mathcal{S}$, and a predefined error level α to quantify the total predictive uncertainty geometrically by generating a prediction set and ensures an upper bound on the errors made by the ML framework.

Aleatoric (or data uncertainty). In a nutshell, data uncertainty relates to the inherent stochasticity in the data-generating process, while model uncertainty represents the lack of knowledge about the best model-unknown data generating process. Gawlikowski et al. (2023) presents a comprehensive overview of uncertainty in DL, further pointing out that epistemic uncertainty is reducible, while aleatoric is generally irreducible. They demonstrate metrics to measure uncertainty and provide various approaches to address uncertainty in the context of real-world applications. Hüllermeier and Waegeman (2019) offers a detailed explanation of uncertainty sources in supervised learning settings and provides a classification of various ML-based methods for dealing with uncertainty and explores set-valued prediction approaches, a further class of methods to estimate uncertainty.

One of the most prominent algorithms for this class is Conformal Prediction (CP) (Vladimir Vovk, 2005; Papadopoulos et al., 2002; Lei and Wasserman, 2014). While it has had theoretical foundations since the early 2000s a more comprehensive treatment has been provided around 2005, CP has recently gained attention in DL (Angelopoulos and Bates, 2021). The fundamental concept is to transform a ML predictive model referred to in the literature as the *Underlying algorithm* into a *Set* predictive algorithm that generates prediction sets for unseen samples or as referred to in the literature, to generate a *Hedged prediction*. A prediction set is defined as a set of labels that guarantees to cover the reference value with a predefined error (significance) level in a classification problem and an interval in the case of regression tasks. In other words, the prediction set/interval ensures coverage of the true reference value with an upper bound for errors created by the ML system while representing the underlying uncertainty geometrically by controlling the set size.

In this paper, we explore and compare conformal prediction with other uncertainty quantification techniques across diverse tasks, scenarios, and datasets within the agricultural context. We begin by outlining the main research gaps in uncertainty quantification for machine learning in agriculture in Section 2. We critically assess CP's strengths and limitations compared to alternative methods, focusing on their ability to convey prediction uncertainty from both qualitative and quantitative perspectives.

We conduct two experiments to highlight the characteristics of CP: first, an ablation study examining the impact of CP parameters on

performance; and second, a comparative study evaluating CP against deep ensembles and softmax techniques in real-world scenarios. Our main contributions are as follows:

- **Macro and Micro Analysis of Conformal Prediction:** We explore conformal prediction at both broad and granular levels through an ablation study, uncovering its unique characteristics and potential impacts on agricultural tasks.
- **Innovative Case Study for Data and Model Insights:** We demonstrate how CP not only provides calibrated uncertainty estimates but also serves as a diagnostic tool to identify data issues and model limitations. Using domain knowledge, we highlight how CP's tunable parameters can optimize performance in agricultural contexts.
- **Real-World Comparison with Established Methods:** We benchmark CP against deep ensembles and softmax techniques under real-world scenarios, showcasing its advantages and shedding light on its limitations.
- **Guidance for Future Research:** We propose forward-thinking research directions to extend CP's utility across different scales, addressing current limitations and inspiring new applications in agriculture and beyond.
- Vision for Democratizing Reliable Models: We share our vision of how CP can transform the agricultural sector by offering interpretable, low-computational-cost models that empower farmers and practitioners to make informed decisions with confidence.

Being agnostic to the model in the sense that no changes are required for the model architecture and no re-training, distribution-free as no prior assumptions are needed for data generating distribution, and coupled with its finite sample statistical guarantees (Angelopoulos and Bates, 2021) which ensures robust statistical performance for practical applications, its limited application in digital agriculture to date underscores the significant potential it holds. It is, therefore, an important layer that should be integrated into our modern data-driven agricultural systems.

The paper is organized as follows: Section 2 reviews related work, followed by an introduction to ICP in Section 3. Section 4 presents the datasets, and Section 5 outlines the experiments, including an ablation study in Section 5.1 and a comparison of ICP with deep ensembles and softmax under real-world scenarios in Section 5.2. Sections 6 and 7 provide final remarks and discuss broader agricultural implications, while Section 8 concludes the paper. Additional analysis and results are available in the Appendix. The code of the paper will be made publicly available on GitHubrepository.

2. Related work

2.1. Uncertainty in machine learning

Addressing uncertainty in machine learning can be approached in two distinct ways. Firstly, one might assume that uncertainty can be decomposed into model and data uncertainties, allowing for their independent estimation. Under this assumption, the combination of these uncertainties constitutes the total predictive uncertainty (Shaker and Hüllermeier, 2021). However, focusing exclusively on data and model uncertainty may overlook other significant contributors to uncertainty in real-world problems, yet it remains crucial for in-depth analysis of the data and the model, particularly if enhancements are desired.

Alternatively, one might adopt the viewpoint that such a decomposition of uncertainty is questionable (Hüllermeier, 2022). In this case, the focus shifts to assessing the total predictive uncertainty, a critical measure for ensuring model reliability without examining the specific origins of the uncertainty (Angelopoulos and Bates, 2021). It is important to acknowledge that the concept of uncertainty, its origins, and the conventional categorization into model and data uncertainties remain ambiguous in ML, as indicated by Gruber et al. (2023). These observations underscore the presence of additional uncertainty sources, such as missing data and the deployment of ML models in dynamic environments, suggesting that the landscape of uncertainty is more complex than traditionally understood.

Epistemic uncertainty quantification techniques. Epistemic uncertainty arises from various factors such as insufficient data, model misspecification, or inappropriate selection of hyper-parameters. It is important to distinguish between two types of epistemic uncertainty: *Model uncertainty* (model misspecification) and *Approximation uncertainty*. The former is due to inadequate model selection, representing the selection of the *Hypothesis space* $\mathbb{H}$. The latter depends on the size the data and the hyperparameters, reflecting the degree to which the model approximates the *True model* within the predefined hypothesis space[1]

Several techniques have been developed for estimating epistemic uncertainty. Among these, Bayesian Neural Networks (BNN) which is a prominent method (Wang and Yeung, 2020). BNNs aim to compute the posterior distribution of the weights by observing the data, from which probabilistic outputs are derived through sampling. However, this approach becomes computationally intractable for DL models due to its complexity. This situation necessitates the development of methods to approximate the intractable posterior distribution. Consequently, stochastic techniques have been created such as Monte Carlo Dropout (MC-Dropout), Ensemble Learning (EL), and Flipout (Gal and Ghahramani, 2015; Lakshminarayanan et al., 2017; Wen et al., 2018). They generate an approximated posterior predictive distribution, by perturbing the weights in a stochastic way based on the used approach. EL seen as BBN approximation (Hoffmann and Elster, 2021) is the most commonly investigated due to enhancing the predictive performance and the quality of the approximation obtained. It involves training multiple copies of the same architecture with different hyper-parameters and dataset splits.

Despite their potential, the computational demands of these stochastic approaches often limit their practical application in addressing real-world problems. Hence, deterministic uncertainty quantification methods (DUMs) (Mukhoti et al., 2023; van Amersfoort et al., 2020; Liu et al., 2020) emerging as an alternative, these methods are designed to mitigate computational overhead during inference by treating model weights as fixed values. They efficiently estimate epistemic uncertainty in a single forward pass by applying regularization to the model's hidden representations. However, under continuous distributional shifts, the calibration of uncertainty estimates by DUMs is considerably worse than that achieved by stochastic methods (Postels et al., 2021). Regardless of the differences, Lahlou et al. (2023) demonstrates that both deterministic and stochastic methods fail to capture model misspecification leading to uncalibrated uncertainty estimates and introduce Direct Epistemic Uncertainty Prediction (DEUP) as a new technique for obtaining more accurate epistemic uncertainty estimates.

Aleatoric uncertainty quantification techniques. Data uncertainty, which arises from issues such as feature overlap and errors in labeling, can be categorized into two types: *homoscedastic* and *heteroscedastic*. The former implies a constant uncertainty regardless of the input, while the latter suggests that uncertainty varies with each specific input. Evaluating aleatoric uncertainty could be done in regression, by modeling not only the mean but also the variance of the target distribution. For example, by assuming a Gaussian likelihood and minimizing

[1] An assumption is commonly made to set model uncertainty to zero, based on the complexity of addressing this uncertainty and the argument that deep neural networks are sufficiently rich to encompass the correct hypothesis space. Thus, the approximation uncertainty is the only significant factor for epistemic uncertainty (Hüllermeier and Waegeman, 2019). We follow the same assumption in our experiments: epistemic uncertainty $\approx$ approximation uncertainty.

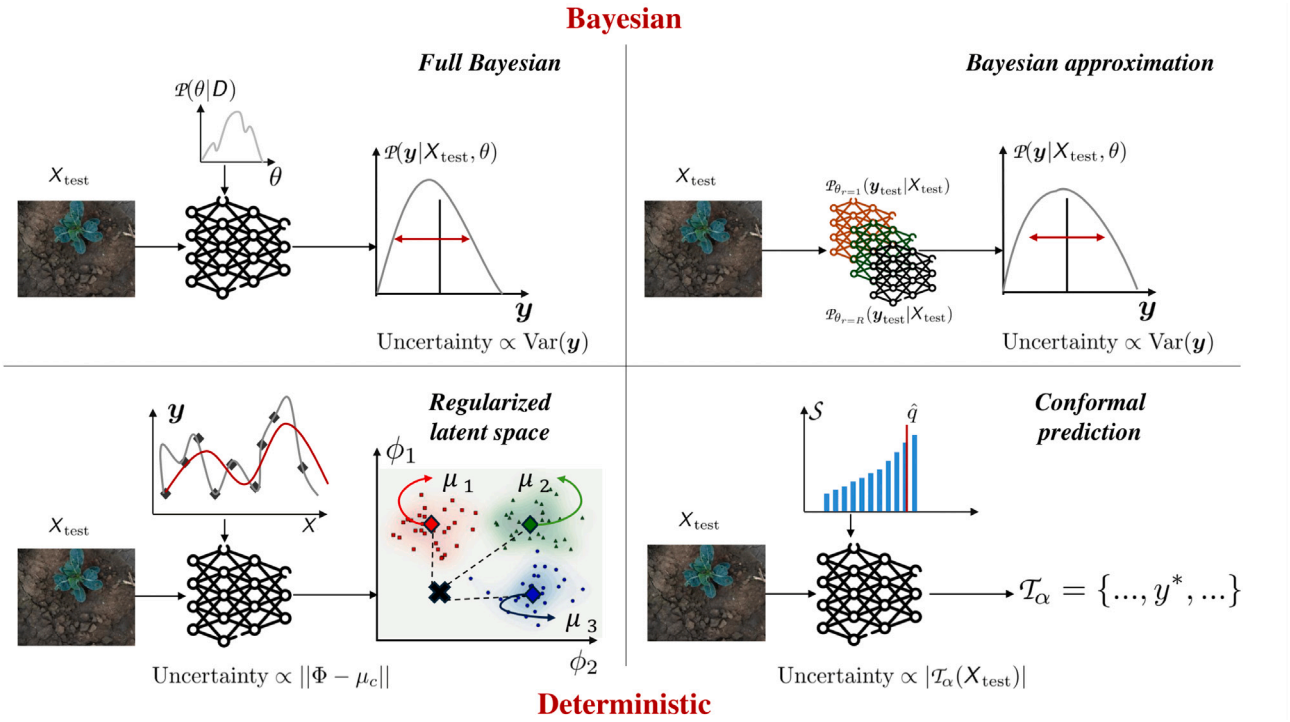


Fig. 2. Uncertainty quantification methods. Uncertainty quantification in machine learning can be broadly categorized into Bayesian and deterministic approaches. Full Bayesian methods estimate uncertainty by exploring the posterior distribution of model parameters, providing probabilistic outputs where uncertainty can be measured using variance or entropy. However, exact Bayesian inference in deep learning is intractable, leading to approximation techniques such as deep ensembles, which uses multiple models to approximate the posterior distribution and quantify uncertainty. To improve real-time feasibility, deterministic methods have been explored as a computationally efficient alternative, often relying on distances in the feature space to estimate uncertainty. Some of these methods implicitly regularize the latent space to enforce smoother representations. However, uncertainty estimation techniques from this family suffer from miscalibration, meaning their uncertainty estimates may not always align with true confidence levels. Conformal prediction provides a post-hoc calibration method that generates prediction sets with guaranteed coverage, where the width of these sets can serve as a proxy for uncertainty.

the corresponding negative log-likelihood, both the mean and the variance can be learned (Nix and Weigend, 1994). Alternatively, quantile regression – which employs the pinball loss function – estimates specific quantiles of the predictive distribution, allowing one to infer uncertainty from the spread of these quantiles (Koenker and Bassett, 1978). In classification, the softmax function converts network logits into a probability distribution over classes, which can serve as a proxy for predictive uncertainty. However Guo et al. (2017a) have critiqued this approach due to its susceptibility to miscalibration, a condition where predicted probabilities misrepresent true likelihoods. It is also prone to overconfidence, where it assigns high probabilities to samples that are incorrectly classified (Lakshminarayanan et al., 2017), thereby distorting the true measure of uncertainty. Moreover, this method falls short in capturing epistemic uncertainty (Gal and Ghahramani, 2015). Multiple approaches have been developed to overcome the previous problems e.g., Kendall and Gal (2017) use a learned loss as a proxy for the heteroscedastic uncertainty, Mukhoti and Gal (2018) explore MC-Dropout as BNNs approximation to estimate aleatoric and epistemic uncertainty for segmentation models. Kirchhof et al. (2023) investigate contrastive encoders to predict latent distributions instead of points for the input. They prove that these distributions recover the correct posteriors of the data-generating process, including its level of aleatoric uncertainty.

The methods described above represent some of the most common techniques for uncertainty quantification in machine learning—particularly in computer vision. However, given the field's vast scope and rapid evolution, a comprehensive review is beyond this paper's focus. Other approaches include generative models, which intuitively estimate both model and data uncertainties; yet, these methods often require large amounts of data, face challenges in selecting appropriate model classes (Hüllermeier and Waegeman, 2019), and can exhibit overconfidence (Nalisnick et al., 2019). Recently, credal sets (Hofman et al., 2024a) have attracted attention for capturing multiple uncertainty components simultaneously, though they tend to be computationally intensive (Hüllermeier and Waegeman, 2019).

We categorize the most commonly used approaches as shown in Fig. 2 into bayesian and deterministic methods. Bayesian methods quantify uncertainty by explicitly or implicitly modeling the posterior distribution over the model parameters. In contrast, deterministic approaches avoid the computational cost of full posterior inference by estimating uncertainty in a single forward pass. In particular, these methods either constrain the learned weights during training to yield more informative representations or employ techniques such as conformal prediction, which measures uncertainty based on distances in the feature space or prediction set widths derived from hypothesis testing.

2.2. Machine learning uncertainty quantification in agriculture

Due to its significance, there have been multiple attempts to integrate uncertainty quantification techniques into the agricultural machine learning framework as shown in Table 1. Wang et al. (2019) use full bayesian approach to estimate uncertainty for subsoil heterogeneity estimations. They specify a prior distribution and use bayes rule to get the posterior as they use model with relatively low number of parameters compared to deep learning models. MC-Dropout and Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) have been used by Hernández and López (2020) as an approximation to full bayesian approach for plant disease detection. Chrispell et al. (2021), Meenken et al. (2021) have investigated full bayesian approaches while using a hybrid modeling approaches which combines machine learning with physical model for real-time crop management and estimating crop sensitivity to weather changes. Multiple studies for different tasks (Chrispell et al., 2021; Xu et al., 2022; Padarian et al.,

Table 1
Uncertainty quantification for machine learning in agriculture.

Author	Task	UQ technique
Wang et al. (2019)	Estimating subsoil heterogeneity	Full Bayesian
Hernández and López (2020)	Plant disease detection	MC-Dropout/SGLD
Meenken et al. (2021)	Real time crop management	Full Bayesian
Chrispell et al. (2021)	Weather-Crop sensitivity	Full Bayesian
Xu et al. (2022)	Precipitation forecasting	MC-Dropout
Padarian et al. (2022)	Soil Spectral Model Uncertainty	MC-Dropout
Werther et al. (2022)	Chlorophyll-a estimation	MC-Dropout
Li et al. (2023)	Soil moisture retrieval	MC-Dropout/Deep Ensembles
Celikkan et al. (2023)	Crops-Weed segmentation	MC-Dropout
Agyeman et al. (2023)	Soil Zinc level estimation	Kriging
Melki et al. (2023)	Crop-Weed classification	Conformal prediction
Farag et al. (2023)	Harvest readiness	Conformal prediction
Kakhani et al. (2024)	Soil organic Carbon estimation	Conformal prediction
Rau et al. (2024)	Soil classification	Last layer Laplace approximation

2022; Werther et al., 2022; Li et al., 2023; Celikkan et al., 2023) have focused on using MC-Dropout or ensembles to estimate data uncertainty or model uncertainty in deep learning models for being computationally efficient (during training) and providing a direct way to measure uncertainty estimates by using metrics such as mean and variance. Conformal prediction for various agricultural applications (Melki et al., 2023; Farag et al., 2023; Kakhani et al., 2024) has gained attention due to its simplicity and robust statistical guarantees for providing calibrated uncertainty estimates. Other technique such Last layer Laplace (LLL) approximation has been used as another approximation to full bayesian (Rau et al., 2024) and kriging by Agyeman et al. (2023) for soil Zinc uncertainty estimation.

Previous studies primarily rely on the Bayesian paradigm for uncertainty estimation, reflecting the field's historical focus on probabilistic modeling. Conformal prediction has gained attention in agriculture as a computationally efficient alternative, offering statistical guarantees with lower overhead than Bayesian methods. However, significant gaps remain at various levels:

1. Full bayesian approaches are computationally expensive during both training and testing, require careful selection of prior distributions, and often suffer from miscalibration, leading to unreliable uncertainty estimates (Papamarkou et al., 2024).
2. Bayesian approximations although they reduce training costs, they still impose significant real-time computational burdens and are highly model-specific. For instance, in deep ensembles, the model architecture and data splitting impact estimates, while in MC-Dropout, the placement of dropout layers and the dropout rate are critical hyperparameters (Verdoja and Kyrki, 2020).
3. Despite its popularity, MC-Dropout exhibits miscalibration on in-distribution samples and tends to be overconfident on out-of-distribution data (Lakshminarayanan et al., 2017), which can be problematic for high-stakes applications.
4. While effective for providing calibrated uncertainty insights on in-distribution data, its performance under distributional shifts and OoD detection has not been sufficiently explored—an important consideration in the dynamic environments of agriculture. Moreover, most studies have not thoroughly investigated the impact of conformal prediction's parameters on the resulting uncertainty estimates.
5. Most studies focus on a single UQ technique, one dataset, and one model for a task rather than comparing multiple methods across varied settings. This narrow evaluation scope limits our understanding of the generalizability of UQ methods and impedes the selection of the most appropriate tool for different contexts.

Overall, these observations highlight the critical need for further research in uncertainty quantification for agricultural applications.

3. Inductive conformal prediction (ICP)

A dataset is defined as $\{\boldsymbol{X}_n, \boldsymbol{y}_n\} \in \boldsymbol{D}$, $n = 1, \ldots, N$, with $\boldsymbol{X}_n \in \mathbb{R}^{H \times W \times B}$ representing a B-dimensional image and $\boldsymbol{y}_n \in \mathbb{R} = [1, \ldots, L]$ denoting the class labels with L being the total number of classes. ICP[2] transforms a conventional predictive algorithm into a so-called conformalized model by generating a prediction sets that are guaranteed to cover the reference value y^* with a pre-defined probability level $(1-\alpha)\%$ on average for a new test sample $\boldsymbol{X}_{\text{test}}$.

Conformalization demonstrated by Fig. 3 goes as follows:

1. A notion of uncertainty such as softmax outputs in classification is selected, we have L classes represented by probability vector $\boldsymbol{p}$:

$$\sigma(\mathcal{F}(\boldsymbol{X})) = \boldsymbol{p} = [p^1, \ldots, p^L], \tag{1}$$

   where $\sigma(\mathcal{F}(\boldsymbol{X}))$ is the softmax output for image $\boldsymbol{X}$ after feature extraction by the DL model $\mathcal{F}$ and generating the logits $\mathcal{F}(\boldsymbol{X})$.
2. We calibrate the model using new unseen i.i.d. samples $\{\boldsymbol{X}_j, \boldsymbol{y}_j\} \in \boldsymbol{C}, j = 1, \ldots, J$, from the same training set distribution by obtaining the softmax outputs for each sample and applying a proper *Scoring* function (Gneiting and Raftery, 2007) $\mathcal{S}(\boldsymbol{X}_m, \boldsymbol{y}_m)$, with a main role to assign values for the prediction errors. It needs to be selected as a hyper-parameter (e.g., hedge loss). The scoring function in this case is defined in the range $[0, 1]$ as:

$$\mathcal{S}(\boldsymbol{X}_m, \boldsymbol{y}_m) = 1 - p_{j,y^*}. \tag{2}$$

   Where p_{j,y^*} is the probability corresponding to the reference value y^* for image j from the calibration set. It is used as a *conformity measure*, to be aligned with the literature we refer to Eq. (2) as *non-conformity measure*. A suitable choice is important and careful design characteristics such as ranking the prediction errors when the model is applied to inputs should be considered (Angelopoulos and Bates, 2021). The so-called non-conformal scores represent the predictive uncertainty, where a large score means less confidence as the original softmax value is low and vice versa.
3. The empirical quantile of the non-conformal scores, denoted as $\hat{q}$, is calculated based on a pre-defined error level α and the number of calibration samples J is used to correct the value due to the finite number of data points:

$$\hat{q} = \frac{\lceil (J+1)(1-\alpha) \rceil}{J} \tag{3}$$

[2] In this paper, we distinguish between two primary types of conformal prediction: Transductive Conformal Prediction (TCP) and inductive conformal prediction, also known as *Split* conformal prediction. We focus on ICP due to its lower computational overhead of $\mathcal{O}(1/n)$ while preserving many of the characteristics of TCP. However, it is important to note that ICP typically exhibits slightly less statistical power compared to TCP, a consequence of splitting the data.

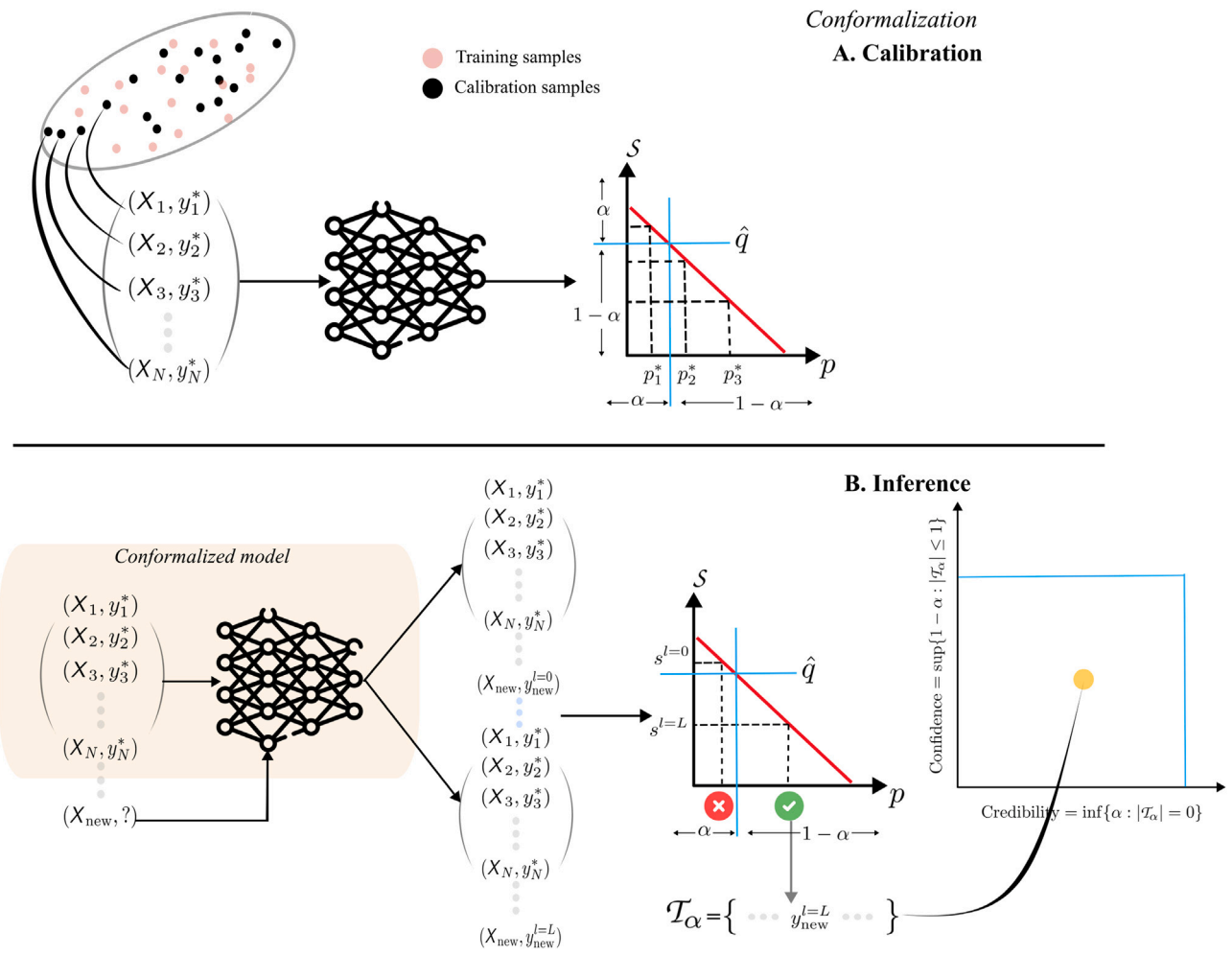


Fig. 3. Inductive conformal prediction framework. Consists of two main stages: calibration and inference. In the calibration stage, we select calibration data from the same distribution as the training dataset and train our model. We then choose an appropriate scoring function, such as hinge loss, and compute a score for each calibration data point. Given a predefined confidence level α (e.g., 95%), we determine the empirical quantile $\hat{q}$. During inference, we use the conformalized model to evaluate a new test point. Each potential label is assigned to the test point, and hypothesis testing is performed under the assumption of exchangeability. If the computed score falls below $\hat{q}$, the label is included in the prediction set $\mathcal{T}_\alpha$; otherwise, it is rejected. Finally, conformal prediction provides two key metrics to assess the reliability of predictions: *confidence*, which measures how likely the assigned label is correct, and *credibility*, which evaluates the ambiguity of the test sample compared to the calibration set.

4. Finally, for a new test sample X_{test}, we apply the scoring function as a measure of the similarity, typicality, or conformity to the scores' empirical distribution formed by the calibration points. We include the label $\boldsymbol{y}^l_{\text{test}}$ if its score falls below $\hat{q}$ to generate the prediction sets $\mathcal{T}$:

$$\mathcal{T}(X_{\text{test}}) = \{\boldsymbol{y}^l_{\text{test}} : \mathcal{S}(\sigma(\mathcal{F}(X_{\text{test}}))) \leq \hat{q}\}. \tag{4}$$

5. The prediction sets for the new unseen samples $\mathcal{T}$ are guaranteed to cover the reference value y^*_{test} by following the probability constraints below:

$$1-\alpha \leq \mathrm{P}(y^*_{\text{test}} \in \mathcal{T}(X_{\text{test}})) \leq 1-\alpha+\frac{1}{J+1}. \tag{5}$$

The probability highlighted in Eq. (5) indicates that on the long run, the obtained sets will include the correct label on average, with a probability of $(1-\alpha)\%$. However, it is not possible to ensure adaptivity (Eq. (5) does not hold for each new sample). Known as conditional coverage, which is a stronger property compared to the marginal coverage in Eq. (5), defined as:

$$\mathrm{P}(y^*_{\text{test}} \in \mathcal{T}(X_{\text{test}}) | X_{\text{test}}) \geq 1-\alpha \tag{6}$$

In most cases, marginal coverage in Eq. (5) is achieved (Angelopoulos and Bates, 2021), while Eq. (6) can only be approximated (Vovk, 2012a).

4. Datasets

Deep Nutrient Deficiency for Sugar Beet (DND-SB) dataset[3] (Yi et al., 2020) is an open source dataset has been created to study how well nutrient deficiency symptoms can be recognized in RGB images of sugar beets. It contains 5648 images of sugar beet plants with seven classes representing the type of deficiency as follows: unfertilized, _PKCa, N_KCa, NP_Ca, NPK_, NPKCa, NPKCa+m+s. Where Nitrogen is represented by (N), Phosphorus by (P), and Potassium by (K). (Ca) refers to liming, _ stands for the omission of the corresponding nutrient treatment or liming, and m+s indicates application of supplementary mineral fertilizer and manure. Hence, class 1 in Fig. 4 refers to full fertilization, class 5 as an example had no liming and supplementary materials and class 2 unfertilized.

GrowliFlower dataset[4] (Kierdorf et al., 2023) is an open-source dataset acquired in 2020 and 2021. It contains 14 K samples of cauliflower plants with multiple modalities available, including RGB and multi-spectral images. The dataset also includes a comprehensive range of annotations, making it useful for computer vision tasks

[3] https://github.com/jh-yi/DND-SB

[4] https://phenoroam.phenorob.de/geonetwork/srv/eng/catalog.search#/metadata/cb328232-31f5-4b84-a929-8e1ee551d66a

Fig. 4. DND-SB -*ZR* stands for *Zuckerrübe* (sugar beet in German)- plants with light green on the left and GrowliFlower on the bottom right with dark green. We have seven classes in the first one which is a representation of the nutrition deficiency detection task. In contrast, in the other dataset, we have a binary classification problem to decide on plants' harvest readiness. For more details check Section 4.

Table 2
Dataset Splits Summary.

Dataset Split	DND-SB	GrowliFlower
Training (I)	2976	6224
Validation and Calibration ($G = C$)	992	196
Calibration (A)	650	100
Validation (B)	342	96
Testing (E)	1632	194

such as classification and segmentation. This paper considers harvest readiness as a binary classification problem, where class 0 indicates not ready for harvest class. We choose images of four acquisition days (23.08.2021, 25.08.2021, 30.08.2021, 03.09.2021) for which the information about harvest-readiness is available. Samples from both datasets are represented in Fig. 4.

5. Experiments

5.1. Experiment 1: Comprehensive evaluation of conformal prediction across models, datasets, and ablation studies

This experiment examines ICP's coverage validity and how it represents uncertainty through prediction set sizes. It explores the trade-off between precision and coverage in tasks like cauliflower harvesting and sugar beet nutrition deficiency detection. Higher precision can reduce coverage, while a more conservative approach maintains coverage but results in larger prediction sets. Balancing precision and coverage is essential for effective decision support, and the quality of both the data and model significantly impacts the size of the prediction sets produced by ICP.

5.1.1. Experimental setup

For both datasets, The total samples D is divided into four subsets: the training set I, the validation set G, the calibration set C, and the testing set E.[5] To evaluate the performance of ICP, we first split the original validation set G into two subsets A and B where $G = A \cup B$. A allocates samples for calibration and samples to evaluate ICP's performance are included in B and the final evaluation of ICP is made on the entire testing set E.

[5] Although some works use identical sets G and C, practically we should use separate sets to avoid model exposure to data seen previously during validation in case of hyper-parameter tuning. However, due to the lack of data we follow the same approach, the entire validation set G is utilized as a final calibration set C ($G = C$).

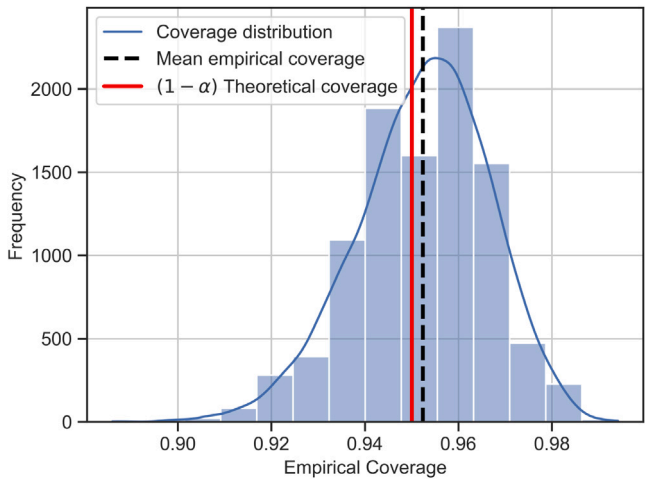


Fig. 5. Empirical Coverage Distribution of ICP. The figure shows the convergence of ViT-B/16 on the DND-SB dataset which demonstrates benign performance of ICP.

Table 3
Conformal prediction settings.

Setting	DND-SB	GrowliFlower
Calibration Samples (C)	992	196
Coverage $(1-\alpha)$ (%)	90.0	80.0
Coverage Error (ϵ) (%)	±1.60	±3.60
Coverage Probability $(1-\delta)$ (%)	90.0	90.0

All images are pre-processed according to the ImageNet (Krizhevsky et al., 2017) pre-processing pipeline. First, they are resized to 256×256 pixels using bi-linear interpolation, then centrally cropped to 224×224 pixels. The pixel values are rescaled to a range between 0 and 1. Finally, they are normalized using mean values of $[0.485, 0.456, 0.406]$ and standard deviations of $[0.229, 0.224, 0.225]$. DND-SB dataset is additionally pre-processed by augmentation using random vertical, horizontal flipping, and rotation by 45°, Random rotation is applied as an augmentation technique to enrich the variability of both datasets to avoid overfitting. Both datasets are split according to Table 2, and we ensure that classes' distribution remains the same across splits by using stratified splitting.

ResNet18 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021) are used and initialized by ImageNet weights for transfer learning, then changing the classification head on top of each model's encoder with the proper number of output classes for each dataset. Both are completely unfrozen and fine tuned on both datasets by using cross entropy loss as a loss function, with learning rate 0.0001 of Adam (Kingma and Ba, 2014) optimizer, and batch size 32 with 500 epochs of training while using early stopping technique with 50 epochs after which training is stopped if no enhancement for validation loss occur.

For ICP, as mentioned in Angelopoulos and Bates (2021) 1000 calibration samples are proper number for most of the applications, and they provide robust method that we can use to calculate the number needed for to achieve $(1-\alpha)\%$ coverage probability with error ϵ and by probability $(1-\delta)\%$ of achieving the desired coverage, Vovk et al. (2022) show that $(10/\alpha)$ points are needed to ensure stable performance. Table 3 presents ICP settings used on our experiment. Finally the NVIDIA RTX A4500 with 16 GB of memory was employed as our hardware accelerator.

5.1.2. Coverage validity and prediction set types in conformal prediction

Results. To illustrate the flexibility of inductive conformal prediction, we apply it to two different models across two distinct datasets, each associated with different tasks. Both models achieved approximately the desired coverage (80% for GrowliFlower and 90% for DND-SB) on both tasks, as demonstrated in Table 3. The mean empirical coverage is calculated as the mean value obtained by randomly splitting the calibration and testing data 10^4 times. Furthermore, as mentioned in Angelopoulos and Bates (2021), ICP's proper implementation is achieved when the mean empirical coverage closely approaches the theoretical threshold and the distribution of coverage values is centered around it for multiple random splits, as shown in Fig. 5.

Table 4
Prediction set types for conformalized ViT-B/16 evaluated on the test set $\boldsymbol{E}$. The upper section shows performance on the DND-SB dataset, and the lower section shows results for GrowliFlower.

Dataset	Metric	Singleton Sets	Non-Singleton Sets
DND-SB	Total	1555	77
	Correct	1465 *(CCPS)*	33 *(CUPS)*
	Incorrect	90	44
GrowliFlower	Total	180	14
	Correct	138 *(CCPS)*	7 *(CUPS)*
	Incorrect	42	7

Table 5
Performance metrics. The upper section shows results for the DND-SB dataset, and the lower section shows results for GrowliFlower.

–	ViT-B/16	ResNet-18
DND-SB		
Acc_{test}	91.8%	92.2%
Empirical Coverage	0.9013	0.9012
GrowliFlower		
Acc_{test}	74.7%	71.1%
Empirical Coverage	0.8070	0.8073

ICP generates prediction sets, in case of classification the number of combinations is 2^L. For GrowLiFlower we have four combinations $\{\phi\}$ (empty set), $\{0\}$, $\{1\}$, and $\{0,1\}$. While for DND-SB we have 128 combinations including the $\{\phi\}$ set. We refer to the sets with only one element as *Singleton* sets, and the remaining types are *Non-singleton*. We have three generic clusters for the outputs:

- *Correct confident prediction set (CCPS)*
  $\leftarrowtail (\hat{y} = y^* \wedge y^* \in \mathcal{T} \wedge |\mathcal{T}| = 1)$
- *Correct uncertain prediction set (CUPS)*
  $\leftarrowtail (\hat{y} = y^* \wedge y^* \in \mathcal{T} \wedge |\mathcal{T}| \rangle 1)$
- *Incorrect prediction*
  $\leftarrowtail (\hat{y} \neq y^* \wedge y^* \notin \mathcal{T})$

Where the first one refers to the ability to ensure coverage while representing high confidence regarding the output demonstrated by the cardinality (size) of the set $|\mathcal{T}| = 1$, the second type ensures coverage but with uncertainty $|\mathcal{T}| \rangle 1$, and for the final one our prediction is wrong. The size of the set represents the uncertainty of the prediction, hence the correct confident predictions are singleton sets, while correct uncertain predictions are non-singleton sets covering the reference value or empty sets with no labels. Fig. 6 shows the three prediction set types that can be produced by ICP.

Table 4 shows the performance of the conformalized ViT-B/16 on both datasets, we get total singleton sets of 1555 and 77 empty sets for DND-SB at the pre-defined coverage error $\alpha = 0.1$, with 1465 confident correct, 33 uncertain correct, and 134 incorrect predictions. For GrowliFlower we have 138 singleton and 7 double element sets at $\alpha = 0.2$.

Discussion. *Coverage* refers to the ability of the prediction set to include the reference value and *Miscoverage* is the opposite. By selecting α error level, the coverage is explained as $(1-\alpha)\%$ of the produced sets for the new unseen samples will include the reference value. For CP the coverage is *Valid*, to highlight the advantage of such characteristic we calculate the percentage of incorrect prediction sets or sets that miscovered the reference value. From Table 5, we have $\frac{134}{1632} = 8.21\%$, and

Fig. 6. Qualitative results for conformalized ViT-B/16 on DND-SB dataset. We present three cases: an easy case where the model confidently predicts the correct class with a tight set (green), a difficult case where the prediction is correct but uncertain, resulting in a wider set covering other candidates (yellow), and an incorrect case where the model misses the correct class (orange).

mis-coverage theoretical limit is $\alpha = 10\%$ for GrowliFlower. This result is only for one run, applying the law of large numbers by randomly splitting the calibration and testing data 10^4 times, the mean empirical coverage is around 90.13%, and mis-coverage is at most 10.0%.

CP is *always valid* mathematically, as mentioned by Vovk et al. (2022). This means that over a long run, the percentage of empirical errors (when applying the system in real world) at the pre-defined error level α will be at most α, as represented by the previously calculated error rate. This concept parallels calibration in machine learning, where a well-calibrated model produces probabilities that accurately reflect the true likelihood of the predicted event. Therefore, CP validity is akin to calibration because the observed probability of errors – specifically, the mis-coverage of the reference value in successive predictions – closely matches the theoretical error rate α or the true likelihood.

We can refer to the valid marginal coverage guaranteed by CP in Eq. (5)as *marginal calibration* (Gneiting et al., 2007), and it is defined as follows:

$$\overline{\mathcal{Q}} = \lim_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{Q}_t(X)\right), \overline{\mathcal{W}} = \lim_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{W}_t(X)\right), \quad \overline{\mathcal{Q}}(X) = \overline{\mathcal{W}}(X), \quad \forall X. \tag{7}$$

Where $\mathcal{Q}$ represents the Cumulative Distribution Function (CDF) of the true data-generating distribution, $\mathcal{W}$ denotes the estimated or the empirical CDF, and T signifies the number of steps or instances. To achieve marginal calibration as indicated in Eq. (7), $\bar{\mathcal{Q}} = \bar{\mathcal{W}}$ must hold for all values of x.

From a theoretical perspective, $\mathcal{Q}$ can be shown as the true CDF of the nonconformity scores, which is unknown in real life. $\mathcal{W}$ refers to the empirical CDF. By replacing $\bar{\mathcal{Q}}$ with the theoretical error rate α, and $\bar{\mathcal{W}}$ with the mean empirical coverage derived from multiple iterations of random data splitting between calibration and testing phases, it can be verified that CP is marginally calibrated. In other words, CP is valid and ensures marginal coverage. Fig. 5 shows marginal calibration of CP, as with more steps we notice that the empirical mean coverage is converging to the theoretical value.

Marginally calibrated CP is probabilistically calibrated (Gneiting et al., 2007) and can be verified as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{Q}_t(\mathcal{W}_t^{-1}(p)) \to p, \forall p \in [0,1]. \tag{8}$$

$\mathcal{W}^{-1}$ is the quantile function which is the inverse of empirical CDF, and the true CDF $\mathcal{Q}$ is evaluated at quantile value generated by $\mathcal{W}^{-1}$ at the probability level p. Replacing p with $\mathcal{W}_t(x)$ makes $\mathcal{Q}_t$ be evaluated on x, similar to the expression in Eq. (7), hence CP is marginally and probabilistically calibrated. The validity/calibration ensured by conformal prediction is distinct as it is applicable to real-world finite datasets which allows for robust statistical error boundaries empirically and not only theoretically dependent.

For cauliflower harvesting this illustrates that on average, 80% of the prediction sets generated will have the true decision, validity at this situation ensures reliability of the system as we know before hand that the percentage of error accepted is 20%. Furthermore, the type/size of the prediction set reflects how ICP handles the underlying uncertainties propagated from model, data, or both. 138 CCPS cover the correct value with high confidence regarding the label being added to the set, while the 7 double element sets ensure coverage but with low uncertainty for which the ML system must abstain and human close investigation is required. The enhancements in decision support is significant as ICP represents uncertainty in an interpretable and calibrated way with a triggered signal that the data and/or the model need to be checked.

For DND-SB, the 1465 CCPS show that we are confident regarding the deficiency type which enhances the decision regarding the treatment needed allowing no harm to the soil due to excessive unneeded fertilization. In contrast to GrowliFlower, at this error level 10%, we have 33 empty prediction sets representing highly uncertain predictions for which human intervention is required to decide for the type of deficiency which is a safety feature by conformal prediction. For this situation, the system is prevented to decide for the fertilization type leading to no damage for the plants and soil.

In the next section, we analyze comprehensively the effect of ICP's parameters on its outputs for ViT-B/16 on both datasets. Starting by checking conditional coverage, the effect of changing α, the number of calibration points C, and the scoring function $\mathcal{S}$.

5.1.3. Analyzing conditional coverage: Impact of alpha, calibration samples, and scoring function on prediction set validity

Conditional coverage assessment. We define coverage as being valid/calibrated marginally, meaning it is averaged across all samples and classes in our datasets as the quantile value $\hat{q}$ is calculated based on the collective information in the calibration points without being based or conditioned on specific group or class. However, we typically strive for prediction sets that are customized for each individual sample, in other words we aim for the validity to be conditional depending on the information related only to this specific sample to ensure higher reliability with tailored prediction set. The proposed metrics

Table 6
Conditional coverage assessment of ViT-B/16.

Metric	DND-SB	GrowliFlower
FSC	0.87 (class 2)	0.78 (class 0)
SSC	0.94 ($\lvert\mathcal{T}\rvert = 1$)	0.76 ($\lvert\mathcal{T}\rvert = 1$)

by Angelopoulos and Bates (2021) are used to assess how close we are to the theoretical limit $1-\alpha$ as an approximation for conditional coverage guarantee. The Feature-Stratified Coverage (FSC), and Size-Stratified Coverage (SSC) metrics use the same equation shown below but the application is different:

$$\text{FSC} = \text{SSC}: \min_{g\in\{1,\dots,G\}} \left(\frac{1}{|U_g|} \sum_{i\in U_g} 1\{y_i^* \in \mathcal{T}(X_i)\} \right). \tag{9}$$

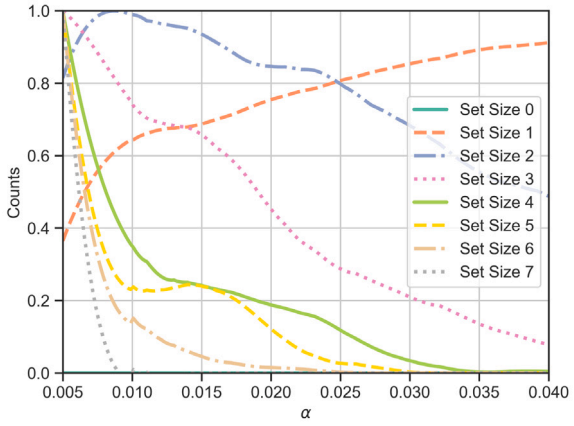
Where G is the number of groups, $|U_g|$ represents the total number of data points in each group g. Index i identifies each data point within the group. The indicator function 1 evaluates to true if the reference value y^* is included in the prediction set $\mathcal{T}$, and false otherwise. For FSC, features can be discrete or continuous and in our case we form different groups based on the classes available for each dataset as images related to same class share similar features. SSC, is forming groups based on the size of the prediction set or in other words its cardinality. If the minimum value across all groups is $1-\alpha$ for both metrics, this indicates that conditional coverage is achieved approximately. A greater divergence from it suggests a larger deviation from achieving this goal.

For GrowliFlower, two groups are established due to the presence of two output types: single and double element sets. Similarly, DND-SB forms two groups, corresponding to singleton and empty sets, in line with the predefined α for SSC. In the case of FSC, outputs are clustered according to the number of classes in each dataset, resulting in two groups for GrowliFlower and seven for DND-SB. Table 6 shows the least empirical coverage for a group among different clusters, although conditional coverage is approximated for both datasets, it is not fully achieved as indicated by Eq. (6).

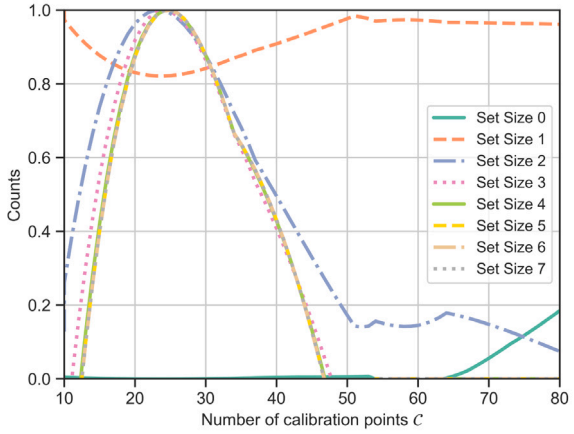
The metrics give intuition when the set is not empty and the correct label is included, we consider other cases such as ICP settings for DND-SB dataset. Empty sets means rejecting all classes at a predefined error rate implies that no label is assigned for a given input. This is considered advantageous when calculating the SSC metric, as it reflects the framework's ability to withhold output when confidence is low. Similarly, when the prediction is incorrect, i.e., $\hat{y} \neq y^*$, having an empty prediction set is beneficial as it prevents the framework from delivering an erroneous output which is the case. In GrowliFlower settings, double element sets indicate high uncertainty regarding the prediction. When $\hat{y} = y^*$ and a non-singleton set is present, this situation reflects a correct yet uncertain output, which is positively regarded in the SSC metric calculation. If $\hat{y} \neq y^*$, the framework adopts a conservative approach by including all possible labels to avoid producing incorrect outputs.

Conditional coverage guarantee ensures that ICP is covering each group equally so no group suffers from under/over coverage which is approximately helps to ensure reliability for each new instant. In case of GrowliFlower, class 0 is under covered by 2.5% meaning that in real life the prediction sets generated would miss the reference value for class 0 by more than 2.5%.[6] As a consequence the decision regarding non-harvesting is less reliable and is not calibrated with our pre-defined value, at this situation we would harvest plants not ready for harvesting which leads to loss of yield. In DND-SB under covered class 2 by 3.33% leads to missing the liming deficiency which may damage the soil and leads to improper treatment.

[6] The values presented in the Table 6 are only for one split between calibration and testing data.



(a) **Effect of changing α on the prediction set cardinality**. The number of counts per set type is normalized.



(b) **Effect of changing the size of the calibration set C on the prediction set cardinality**. The number of counts per set type is normalized.

Fig. 7. Comparison of the effects of changing α and the size of calibration set C on the prediction set.

SSC shows that singleton sets for DND-SB over cover and in the case of GrowliFlower under covers. The former case indicates that ICP is conservative to control the error, however this lead to not being *Efficient/Precise/Optimal*, while for the latter case ICP is precise instead of controlling the error. We show in the next analysis the difference between both by only analyzing ViT-B/16 on DND-SB as it has more classes so the results would be sufficiently significant.

Effect of changing α and the size of calibration dataset C on the prediction sets. As shown in Figs. 7(a) and 7(b) that increasing the pre-defined error rate or in other words reducing the desired coverage will reduce the multi-element sets and increases the empty sets. The similar effect is observed for the number of the calibration points.

During testing, the scoring function estimates a score for each class label and compares it with the calibration empirical distribution under the assumption of data *exchangeability* which is a weaker assumption compared to i.i.d., as it represents the permutation invariance of the data/score's joint distribution. This comparison essentially is a form of hypothesis testing (Neyman and Pearson, 1933), where we test the hypothesis of exchangeability of a new unseen input to the calibration points (Vovk et al., 2022). Any score larger than $\hat{q}$ rejects the null hypothesis, otherwise considered a candidate for the respective label, and the label is added to the prediction set.

Reducing α increases $\hat{q}$ value assuming we have the same number of the calibration points. From a mathematical point of view, as the quantile splits the empirical distribution into two regions -typical and atypical- the more the increase the easier for the new sample scores not to reject the hypothesis of exchangeability and to be added to the prediction set which gives rise to wide sets at the left region of Fig. 7.

Intuitively speaking, the implicit effect of increasing $\hat{q}$ could be represented as adding more samples to the conformal region. The new sample now is tested to a larger group of data points which makes it easier to ensure similarity/typicality with the new group and difficult to reject the hypothesis, and the framework adds multiple labels inside the prediction set. While increasing α reduces the number of samples that are considered typical, and make it harder for the new sample to confirm typicality, as it is tested to a smaller group of points which needs from the new testing point to contain features closer to the new smaller group thus the model start to reject some labels' scores.

The framework in the first condition is more conservative and less precise, and in the second situation is the opposite. With increasing the error rate, the model rejects all of the scores and generates empty sets cause no label could make the new point exchangeable in other words the new sample point lacks the features that make it similar to the small group.

Increasing the number of calibration points C for the same predefined error level increases $\hat{q}$. This makes it easier for the new sample's scores not to reject the hypothesis, and more labels are added to the prediction set, leading to conservative behavior. Mathematically, if the number of calibration points is only $J = 10$ and $(1-\alpha) = 90.0\%$ the quantile value will be increased severely to 99.0% which makes our system highly conservative.

Intuitively, at higher coverage levels $(1-\alpha)\%$, the less data there is, the more conservative the framework becomes. To achieve this level of coverage, the system should include all labels because the low number of samples creates a low chance for the new point to be typical to each one of them, as a consequence ICP includes all of the labels. The more we increase the number of samples, the better the framework can construct typical clusters within the conformal region to compare the new sample to, forcing the framework to reject more labels' scores.

The variation of α and the number of calibration points C create a trade-off between being precise and being conservative. In other words, being conservative here means the framework is focusing more on validity/calibration, which represents error control. The more conservative the framework, the less error we have, but the less precise we are which is useful when the goal is to avoid overconfident predictions. Conversely, being precise/optimal is a demonstration of efficiency, as in the case of singleton sets, which means we have a higher chance of error in missing the reference value but this is the only set type that is useful for taking a decision immediately without abstaining. Being efficient means being sharp, as sharpness reflects the concentration of the predictive distribution (Gneiting et al., 2007). It is always a trade off as the sharper the ICP, the better it is conditioned on validity restrictions. The behavior of ICP depends on how could we tolerate being precise while controlling the error which relies on the nature of the application.

Effect of scoring function $\mathcal{S}$ on the empirical coverage. The scoring function is one of the important components of conformal prediction, as it is the medium to transform the normal model to a conformalized one. We keep the number of calibration points the same, and change α to 0.02 for the generation of different prediction sets types. Eq. (2) is the hinge loss (Rosasco et al., 2004) and used as a non-conformity measure. We explore different scoring function which is the margin loss, described as follows:

$$\mathcal{S}(\boldsymbol{X}, \boldsymbol{y}) = \max_{\boldsymbol{y}^{l} \neq y^{*}} \boldsymbol{p} - p_{y^{*}}. \tag{10}$$

Where $\boldsymbol{p}$ demonstrates the probability vector from softmax function, $\max_{\boldsymbol{y}^{l} \neq y^{*}}$ is searching for the maximum difference in probability values between the reference value's probability $p_{y^{*}}$ and other incorrect class probabilities, in other words searching for the incorrect class with the highest probability. If the value is negative or zero it implies the framework confident regarding the new sample, otherwise it illustrates uncertain situation.

Table 7
Effect of ICP's scoring function $\mathcal{S}$. The cardinality of the prediction set $|\mathcal{T}|$ is shown, and the mean empirical coverage is calculated over 10,000 runs with random splits between calibration and testing.

$\|\mathcal{T}\|$	Metric	Hinge Loss	Margin Loss
1	Cardinality	1206	1394
2	//	241	188
3	//	119	45
4	//	42	5
5	//	22	–
6	//	2	–
–	Empirical Coverage	0.9871	0.9681
–	Mean Empirical Coverage	0.9819	0.9655

Table 7 shows total number of each cardinality for $\mathcal{T}$ when using each loss. We observe efficient or sharp behavior when using margin loss compared to hinge loss represented by the increase in the number of singleton sets and elimination of high order cardinalities. Furthermore, the empirical coverage is reduced by 1.9% as another sign of becoming less conservative. Hinge loss only uses the reference class's probability, without considering how the remaining probability mass is distributed among the incorrect classes. This approach tends to be more conservative to ensure coverage. In contrast, margin loss provides more insights into the distribution of other probabilities, making ICP more precise.

In the next section, we investigate a case study at which ICP can be used to check data issues, comparing models and controlling decision support by including domain knowledge.

5.1.4. Case study

We conduct a focused investigation into its application for diagnosing soil nutrient deficiency in sugar beet crops (DND-SB). Soil nutrient deficiency arises when critical elements essential for plant growth are depleted, compromising agricultural productivity and ecosystem stability. These deficiencies are driven by intensive farming practices, soil erosion, and inadequate land management, causing significant risks to global food security (Food and Agriculture Organization (FAO), 2015). This issue also contributes to broader challenges such as *hidden hunger* (micronutrient malnutrition in humans) (Lowe, 2021) and environmental degradation (e.g., *eutrophication* and soil toxicity) (Mustafa et al., 2023).

To simulate a realistic agricultural scenario, we consider the deployment of a machine learning model on an Unmanned Ground Vehicle (UGV) tasked with detecting soil nutrient deficiency in sugar beet fields. In this scenario, the model provides highly accurate point-wise predictions but lacks confidence indicators, meaning that uncertainty is not explicitly accounted for in the decision-support process. Without a measure of uncertainty, soil treatment strategies risk being suboptimal—either missing critical nutrient deficiencies due to overconfidence in the model's predictions or applying excessive treatments unnecessarily, leading to environmental and economic inefficiencies.

As demonstrated in previous experiments, ICP allows us to transform the base model into a conformalized version, which provides uncertainty-aware predictions while maintaining statistical guarantees. In this context, the tuning parameter (α) plays a crucial role:

- A lower α value ensures a more conservative approach, increasing coverage to avoid missing any necessary treatments.
- A higher α value makes the system more precise, prioritizing only the most critical deficiencies while reducing unnecessary interventions.

Domain knowledge can be leveraged to fine-tune α dynamically, ensuring a balance between precision and coverage based on real-world agricultural conditions. By adjusting this threshold appropriately, ICP-enhanced models can help optimize soil treatment strategies, improving soil health and maximizing crop yield while minimizing waste.

Practically, for DND-SB:

- **Problem in Class 4 (*ZR, NP_Ca*)**: - For both models on the DND-SB dataset, class 4 has the largest average prediction set size. - ViT-B/16 exhibits a higher average prediction set size (1.5784) compared to ResNet-18 (1.5450).[7]
  - **Informed decision**: We have detected a potential absence of potassium, although our results are highly uncertain. Therefore, we will refrain from treating the plant until further investigation of the class 4 samples is completed, thereby reducing the risk of soil mistreatment.
- **Model selection**: - For other classes, ResNet-18 consistently generates wider prediction sets than ViT-B/16.
  - **Informed decision**: We propose two possible explanations:

    1. ResNet-18 is inherently less confident than ViT-B/16.
    2. ViT-B/16 is overconfident compared to ResNet-18.

  To analyze this behavior, we compute the Expected Calibration Error (ECE) (Guo et al., 2017b) and find that ViT-B/16 is miscalibrated by 20.83%, indicating significant overconfidence (see Fig. 8). In contrast, ResNet-18 is better calibrated even without conformal prediction, computationally cheaper, and requires less training data, making it the preferred choice.

Our simulated case study shows that conformal prediction not only produces calibrated uncertainty estimates but also acts as a diagnostic tool for individual data points and model selection—and it can be easily tuned based on domain knowledge. The size and validity of its prediction sets depend strongly on the quality of both the model and the data. With high-quality inputs, conformal prediction yields smaller, more accurate sets; with noisy data or a poorly fitted model, it expands the prediction sets to maintain the desired coverage—a built-in safety feature in real-world applications.

5.1.5. General discussion

In this experiment, we examine conformal prediction's coverage validity, calibration, and its representation of uncertainty via the size of its prediction sets. We also evaluate the impact of its outputs on decision support across various agricultural tasks. We present an ablation study of ICP's parameters, highlighting the trade-off between validity and precision.

We stress the need for conditional guarantees and for a conformal predictor that delivers both valid and precise prediction sets. We then investigate a case study demonstrating how conformal prediction can be used for decision support as well as for diagnosing data quality and model performance.

5.2. Experiment 2: Evaluating inductive conformal prediction versus ensemble learning and softmax for OoD detection and covariate shift

In this experiment, we evaluate ICP in real-world scenarios, focusing on its performance under Out of Distribution (OoD) detection and distributional shifts. We introduce two key metrics for ICP: confidence and credibility. ICP is compared with deep ensembles and softmax outputs, commonly used methods for uncertainty quantification, to highlight their respective advantages and limitations.

[7] The average set size is calculated for correct predictions only.

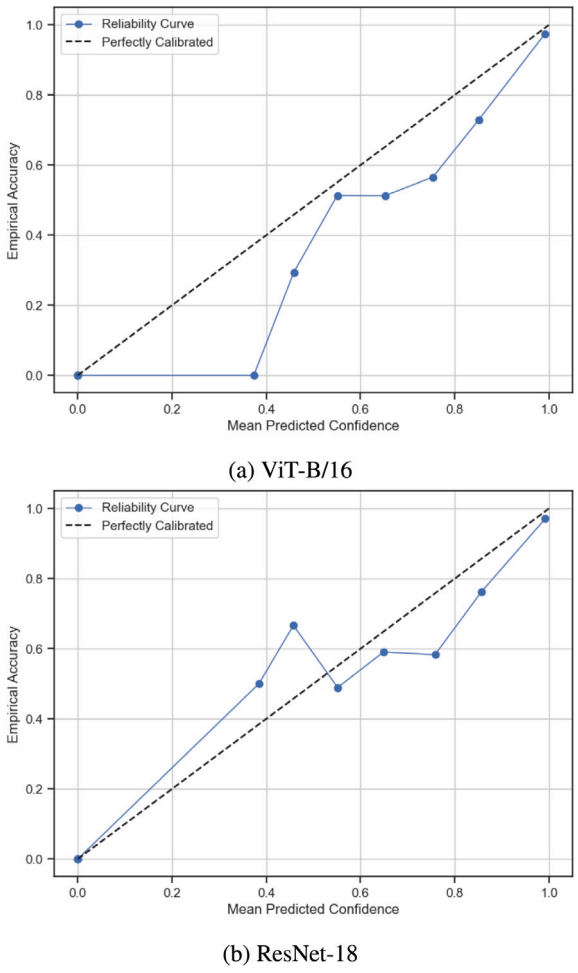


Fig. 8. Reliability plot for different models. As shown in the figure, ViT-B/16 exhibits overconfidence across all confidence ranges compared to ResNet-18, which demonstrates better calibration.

5.2.1. Uncertainty quantification tools

Deep Ensembles. Ensembling is a technique has been used to combine multiple models to enhance predictive performance and reduce over-fitting (variance), described by the following equation:

$$\mathcal{F}_{\text{avg}}(X) = \frac{1}{M}\sum_{r=1}^{M}\mathcal{F}_m(X), \tag{11}$$

where $\mathcal{F}_{\text{avg}}(X)$ is the final model obtained after averaging, $\mathcal{F}_m(X)$ is the mth model, and M is the total number of ensemble members.

Tackling epistemic uncertainty, as mentioned in Section 2, can be achieved using BNNs. We obtain the posterior distribution of the parameters by updating the prior with observed data. The posterior predictive distribution is then generated from the parameters' posterior, and a prediction is made by calculating the expected value (Hüllermeier and Waegeman, 2019).

Deep ensembles could be seen as an approximation to BNNs (Lakshminarayanan et al., 2017), where randomly initialized models combined represents the approximate posterior parameters' distribution $\mathcal{P}(\theta|D)$ for the true posterior $\mathcal{P}^*(\theta|D)$. Hence, the approximate predictive posterior is calculated as follows:

$$\mathcal{P}(\boldsymbol{y}|X) = \sum_{m=1}^{M}\mathcal{P}(\boldsymbol{\theta}_m|D)\mathcal{P}_{\boldsymbol{\theta}_m}(\boldsymbol{y}|X). \tag{12}$$

$\mathcal{P}(y|X)$ is a result of element wise multiplication of $\mathcal{P}(\boldsymbol{\theta}_m|D)$ which is the parameter's posterior and $\mathcal{P}_{\boldsymbol{\theta}_m}(\boldsymbol{y}|X)$ the probabilistic output for each model m. To make a point wise prediction, we search for the class with

the maximum probability at $\mathcal{P}(y|X)$. Eq. (12) is a discretization of the original predictive posterior whose quality is affected by the number of members M.

Softmax outputs. A common approach to assessing uncertainty involves analyzing the probabilistic output of the softmax function. However, as highlighted by Guo et al. (2017a), this method can be unreliable due to calibration issues. Softmax outputs also tend to exhibit overconfidence (Lakshminarayanan et al., 2017), which hampers effective uncertainty detection and quantification. Furthermore, softmax alone cannot capture epistemic uncertainty, which stems from the model's lack of knowledge or limited data (Gal and Ghahramani, 2015).

5.2.2. Evaluation metrics and experimental setup

ICP's metrics. In addition to set size, we introduce two metrics that have been proposed by Gammerman and Vovk (2007). We calculate confidence and credibility for our conformal model while generating a prediction for a new data point to assess the reliability. Confidence is represented as follows:

$$\text{Confidence} = \sup\{1-\alpha : |\mathcal{T}_\alpha| \leq 1\}. \tag{13}$$

Where $\sup\{\}$ is the *Supremum* of the prediction set $\mathcal{T}$ which is the least upper bound for which we can trust the prediction. High confidence indicates that all other labels are unlikely except for the one included in the set. However, confidence alone cannot reveal the sample's ambiguity with respect to the calibration distribution. Credibility is needed to fully understand the strangeness of the sample and it is calculated as:

$$\text{Credibility} = \inf\{\alpha : |\mathcal{T}_\alpha| = 0\}, \tag{14}$$

being the greatest lower limit -*Infmum*. If we have low credibility this signals a problem that needs to be investigated as it measures if this sample is typical or not compared to the calibration samples while confidence measures how accurate we are regarding our sets containing reference value. According to Gammerman and Vovk (2007), if we have confidence close to 100% and credibility more than 5%, the prediction is reliable.

Ensemble's metrics. We follow the same analysis by Shaker and Hüllermeier (2021), Hüllermeier (2022) to calculate the total, aleatoric and epistemic uncertainties for new unseen sample X_{new} based on the *Shannon entropy* $\mathcal{H}(Y)$ decomposition (Shannon, 1948) into *Conditional entropy* and the *Mutual information*. The total predictive uncertainty is estimated as follows:

$$\mathcal{H}_{\text{Total}}(X_{\text{test}}) \approx -\sum_{l=1}^{L} \mathcal{P}(\boldsymbol{y}_l|X_{\text{test}}) \log_2(\mathcal{P}(\boldsymbol{y}_l|X_{\text{test}})). \tag{15}$$

Where $\mathcal{P}(\boldsymbol{y}_l|X_{\text{test}})$ represents the probability of class l derived from the predictive posterior. The minimum value is 0 which reflects total confidence and the maximum is $\log_2(L)$ showing complete uncertainty regarding the new sample. While the aleatoric part is represented by:

$$\mathcal{H}_{\text{Aleatoric}}(X_{\text{test}}) \approx -\sum_{m=1}^{M} \mathcal{P}(\boldsymbol{\theta}_m|D)\times \sum_{l=1}^{L} \mathcal{P}_{\boldsymbol{\theta}_m}(\boldsymbol{y}_l|X_{\text{test}}) \log_2(\mathcal{P}_{\boldsymbol{\theta}_m}(\boldsymbol{y}_l|X_{\text{test}})). \tag{16}$$

The equation shows the weighted sum of the entropy from each model's predictive distribution across the ensemble members M and it is equivalent to the conditional entropy. Epistemic uncertainty, which is equivalent to the mutual information, can be determined as the difference between the total uncertainty and the aleatoric uncertainty:

$$\mathcal{H}_{\text{Epistemic}} \approx \mathcal{H}_{\text{Total}} - \mathcal{H}_{\text{Aleatroic}}. \tag{17}$$

For deep ensembles, we select the number of models to be five which provides proper approximation while reducing the computational cost (Ovadia et al., 2019). The training data I is used to train all replicas, while ensuring random weights' initialization and random augmentation of data during model fitting. ResNet-18, batch size, preprocessing pipeline, and the optimizer from Section 5.1.1 are used with the same settings, and we conduct our study on DND-SB dataset. For the computational demands we switch to NVIDIA A100 GPU with 40 GB of memory as a hardware accelerator.

5.2.3. Out of distribution (OoD) detection

In dynamic real-world scenarios, rare events play a crucial role in determining the confidence in our model's decisions. OoD detection involves identifying samples that are semantically different from the training data, which helps prevent false label assigning of such data (Salehi et al., 2021). This process is crucial for assessing the reliability and safety of our models. An essential method for evaluating the quality of an uncertainty quantification technique is to observe its response to these OoD samples. This response reflects the model's ignorance – expressed as high epistemic uncertainty – about these points. In situations with high uncertainty, abstaining from making predictions and seeking assistance from a human operator should be considered.

The problem of OoD detection can be formulated as follows:

$$\begin{cases} \text{ID}, & \text{if } \psi(X_{n+1}) \geq \tau \\ \text{OoD}, & \text{if } \psi(X_{n+1}) < \tau \end{cases} \tag{18}$$

When X_{n+1} is declared the OoD detector assigns a score ψ to the sample, if it greater than or equals τ it is considered ID, otherwise it is rejected and classified as OoD. For scoring we can use confidence, entropy, or any function under the condition that it is able to differentiate between the classes or to rank them differently. Hence OoD detection can be treated as a binary classification problem, the only difference is that we lack the reference values for the OoD samples.

We conduct an experiment by comparing the three methods trained on the DND-SB dataset, with the GrowliFlower training data serving as out-of-distribution data. The area under (Receiver Operating Characteristics) ROC curve - (AUCROC) (Ling et al., 2003), confidence-rejection and ID/OoD histogram plots are used to articulate the comparative study quantitatively and qualitatively respectively. The number of the OoD sample are 6224 which is the GrowliFlower training set, and for the ID we have 1632 sample which is the testing set from DND-SB dataset.

In our analysis, confidence is used as the scoring approach for softmax and deep ensembles by taking the maximum probability value from the output vector $\mathcal{P}(y = \hat{y}|X) = \max_l \mathcal{P}(y = l|X)$. For ICP, we use the confidence from Section 5.2.2. We set $\tau \in [0, 1]$ as a threshold to assess accuracy (rejection) at each value. Rejection curves are then plotted against the confidence values for OoD samples only. To qualitatively evaluate the separation between ID and OoD samples, we use histograms to visualize the confidence distribution. Finally, for the AUCROC metric, pseudo labels 1 and 0 are assigned to ID and OoD samples, respectively.

Results. As shown in Fig. 9, deep ensembles demonstrate a better separation of ID samples from OoD samples compared to softmax and conformal prediction. While softmax outperforms ICP, both methods lack explicit modeling of the epistemic component, leading to overconfidence on OoD samples. The confidence-rejection plots provide additional insights into each method's sensitivity to OoD samples. The y-axis represents the number of samples classified as ID, while the x-axis shows the threshold τ used to compare the generated scores from the three methods.

In Fig. 10(a), we sweep over the τ parameter, which corresponds to confidence scores. When $\tau \in [0, 0.4]$, all methods classify all OoD samples as in-distribution. Beyond this range, the deep ensemble demonstrates a stronger ability to reject OoD samples compared to the other approaches. For example, at $\tau = 0.6$, the ensemble rejects approximately 2700 samples, while softmax and ICP reject only 420 and 0, respectively. Additionally, the steep slope of the ensemble's rejection

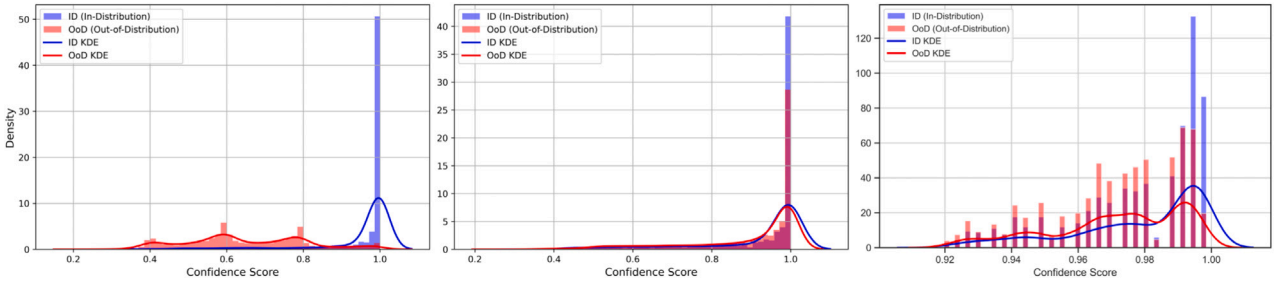
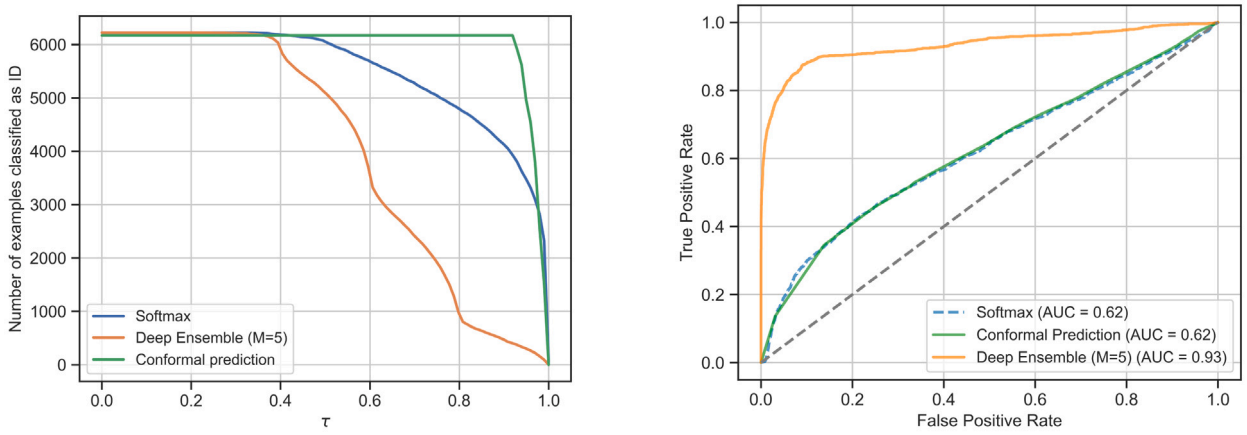


Fig. 9. Qualitative comparison of ID/OoD separability. Deep ensemble on the left is performing better in identifying OoDs compared to softmax in the middle and ICP on the right.



(a) Confidence-rejection curve (b) AUCROC

Fig. 10. Results on GrowliFlower as OoD. Both figures show that deep ensembles outperform softmax and ICP in distinguishing out-of-distribution samples from in-distribution data.

curve indicates higher sensitivity to OoD samples, in contrast to softmax and ICP, which only begin to respond at high confidence levels. This delayed response could lead to severe consequences in high-stakes real-world applications.

AUCROC in Fig. 10(b) evaluates how well each approach separates ID and OoD samples, with the best value being 1 and the worst being 0. The diagonal line represents random guessing, serving as the baseline. Deep ensembles achieve superior performance with an AUC of 0.93, outperforming both softmax and ICP. While softmax and ICP perform better than the baseline, they exhibit limited ability to distinguish OoD samples from in-distribution samples. For ICP, despite the advantage of a sharp drop in coverage to 46.0%, which provides an additional signal for assessing new input samples, its practical use is constrained. This limitation arises because ICP requires estimating empirical coverage over multiple samples, a scenario that is not always feasible in real-world applications.

5.2.4. Calibration under distribution shift

Dataset shift can be rigorously defined, following Moreno-Torres et al. (2012), as a discrepancy where the joint distributions of training and test data are not equal, expressed as $\mathcal{P}_{\text{train}}(X, \boldsymbol{y}) \neq \mathcal{P}_{\text{test}}(X, \boldsymbol{y})$. Multiple types of shifts can occur, including *Prior* probability, *Concept*, and *Covariate* shifts. This paper focuses on the latter, which is formulated as follows:

$$\mathcal{P}_{\text{train}}(\boldsymbol{y}|X) = \mathcal{P}_{\text{test}}(\boldsymbol{y}|X) \quad \text{but} \quad \mathcal{P}_{\text{train}}(X) \neq \mathcal{P}_{\text{test}}(X). \tag{19}$$

The above equation shows that this occurs when the functional relationship between the covariates X and the target $\boldsymbol{y}$ remains unchanged, while the distribution of the input variables differs between the training and testing samples.

We simulate covariate shift scenarios by altering the crop size at the center of the image in the DND-SB test set. We compare the three methods based on accuracy and coverage, prediction set size, and the ratio of coverage to the prediction set size. We set $\alpha = 5.0\%$ to match the original calibration value. In the context of conformal prediction, coverage is interpreted as $(1-\alpha)\%$, meaning that this proportion of the generated prediction sets is expected to contain the true label.

To create a comparable framework for softmax and deep ensembles, we convert them into set predictive algorithms. Using the same α for conformal prediction, we add each sample to the prediction set if its generated probability exceeds the threshold. We then calculate coverage by counting how often the true value appears in the prediction set.

Results. As shown in Fig. 11, performance declines significantly at smaller crop sizes, as all models struggle to identify classes due to the substantial loss of visual information. Increasing the crop size preserves more data context, leading to improved accuracy. Deep ensembles demonstrate better performance under distributional shifts compared to softmax and conformal prediction benefiting from its models' diversity, which enhances predictive performance.

We divided the range into three parts – high, mild, and low covariate intensity – corresponding to the left, middle, and right sections of Fig. 12(a). Under severe shifts, all models struggle to include the true label due to significant loss of visual information. The deep ensemble exhibits higher coverage than softmax and ICP, while ICP performs poorly with a coverage value near 0, meaning almost no prediction set includes the true label. When the shift is reduced all approaches are able to have higher coverage till we reach the original one around the original image size.

Coverage alone is not sufficient to explain the observed behavior; therefore, we also analyze the sizes of the prediction sets in Fig. 12(b)

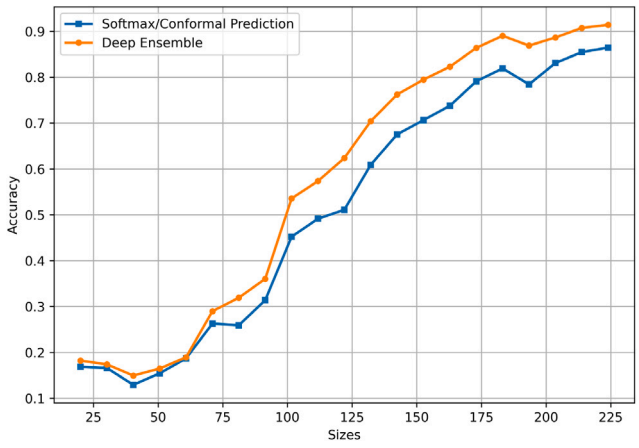


Fig. 11. Accuracy during covariate shift. During severe shifts on the left-hand side, all models perform poorly. However, when the shift is reduced and visual information is restored, the deep ensemble outperforms the others.

generated by each method. While coverage indicates the calibration of the approach, the prediction set size reflects its efficiency or precision. An ideal approach would achieve the pre-defined coverage value with minimal prediction set sizes, as discussed in Section 5.1.3.

Ensembles and softmax generate high cardinality prediction sets under severe shifts because both are less confident, and the probability mass is not concentrated around a specific class. This results in larger prediction set sizes, enabling higher coverage rates. The distinction between softmax and ensembles becomes apparent at the highest shift, around size 25 × 25. Ensembles produce larger prediction set sizes, indicating a better distribution of probabilities in this scenario, while softmax shows overconfidence, reflected by smaller prediction set sizes and lower coverage.

Examining ICP, we observe zero coverage with a very small cardinality at size 25. This occurs because the current data is entirely different from the calibration dataset, violating the cornerstone assumption of exchangeability in conformal prediction. As the shift decreases, coverage improves, and the prediction set size increases due to the higher remaining uncertainty. When the shift is low on the right-hand side, the prediction set size decreases as confidence grows, while coverage continues to increase

In the low shift region, we observe increased coverage across all approaches, accompanied by reduced prediction set sizes due to a better probability mass distribution. Notably, both softmax and ensembles achieve higher coverage at the image size 225 × 225, which is the original size. However, this is misleading, as their wider prediction sets increase the likelihood of including the true label, but the probability distributions are not sharp enough.

For the complete picture we calculate and plot the coverage to prediction set size ratio for each shift value in Fig. 13. As mentioned before the ideal approach ensures coverage (calibration) while being efficient (precise), in other words producing the smaller prediction sets that includes the correct label $(1 - \alpha)$ times. From this figure we can see clearly that at the severe shift region all methods suffer to achieve both.

When the shift is reduced to a mild level, ICP and ensembles outperform softmax, demonstrating a better ability to cover the true label while producing smaller prediction sets. In the low shift region, ICP shows superior performance compared to both ensembles and softmax, achieving smaller prediction set sizes that still include the true label. This makes ICP preferable for making well-informed decisions, as smaller prediction sets (singletons) allow for immediate decision-making without abstention.

5.2.5. General discussion

Overall, our results highlight the advantages of deep ensembles in delivering robust uncertainty estimates, effectively detecting OoD samples, and maintaining strong performance under distributional shifts. While softmax and ICP methods provide baseline performance, their limited capacity to model epistemic uncertainty presents challenges in agricultural scenarios such as crop classification, disease detection, and yield prediction, where novel or changing conditions (e.g., new diseases, weather anomalies, or unseen crop varieties) are common.

As shown in Table 8, ICP, in its basic form, struggles to handle OoD data due to its core assumption of exchangeability between the test point and the calibration samples. When this assumption is violated under such shifts, ICP fails to provide calibrated uncertainty estimates, limiting its practical utility in dynamic agricultural environments. Nevertheless, it demonstrates an acceptable performance compared to the other two methods during low distributional shifts.

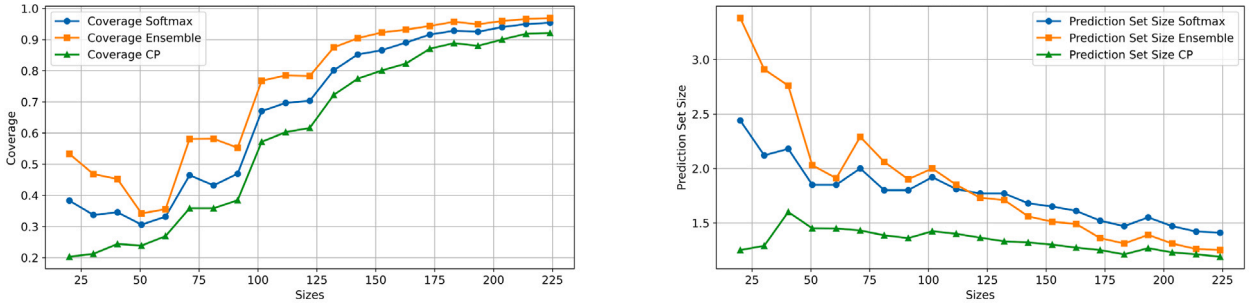
ICP is computationally inexpensive during training and testing, comparable to softmax, and offers the additional advantage of being a post-hoc approach. This reduces the number of hyperparameters to tune compared to deep ensembles while still providing calibrated uncertainty estimates. However, ICP primarily focuses on total predictive uncertainty without distinguishing its source, unlike deep ensembles.

Depending on the application context, different tools may be more suitable. For instance, if real-time performance with calibrated estimates is a priority, conformal prediction is an appropriate choice. Conversely, in applications like active learning (Li et al., 2024), where uncertainty quantification is used to improve model performance, deep ensembles can offer robust performance on OoD data. Softmax remains a simple and quick option, but its limitations must be carefully considered. In conclusion, selecting the appropriate uncertainty quantification method depends on the specific requirements of the agricultural application, balancing the trade-offs between computational cost, robustness to distributional shifts, and the need for calibrated uncertainty estimates.

6. Remarks

While we demonstrate the effectiveness of ICP in uncertainty quantification for agricultural applications, several frontiers for future research remain. The key limitations are as follows:

1. **Labeled calibration set:** Conformal prediction provides statistical guarantees but relies on a labeled calibration dataset. This dependency presents challenges when calibration data is limited in quantity or quality.
   Promising solutions: Cross-conformal prediction (Vovk, 2012b) offers an alternative when calibration samples are scarce. To address the issue of poor-quality calibration sets, merging Bayesian approaches to assess whether calibration points are drawn from the same distribution as the training set could improve robustness. Additionally, Stutz et al. (2024) proposes a modified framework for cases with uncertain ground truth labels.
2. **Exchangeability assumption:** Conformal prediction relies on the assumption that new unseen points are exchangeable with the calibration set. This limits its effectiveness in handling out-of-distribution (OoD) samples and distributional shifts.
   Promising solutions: From a theoretical perspective, Barber et al. (2022) introduces modifications to improve robustness beyond the exchangeability assumption. Adaptive methods explored by Angelopoulos et al. (2024) is promising and introducing conformal prediction as a risk control approach, offering a more flexible framework applicable to diverse tasks while retaining its core guarantees. Cabezas et al. (2025) combines bayesian approaches with conformal prediction to tackle epistemic uncertainty. Novello et al. (2024) investigates the link between OoDs and conformal prediction showing that with modifications to the scoring function, we can tackle such a problem.



(a) Normalized coverage (b) Normalized prediction set size

Fig. 12. Coverage and prediction set size during covariate shift. In both figures, severe shifts lead to reduced coverage and changes in prediction set sizes: softmax and ensemble methods show increased set sizes, while conformal prediction shows a reduction.

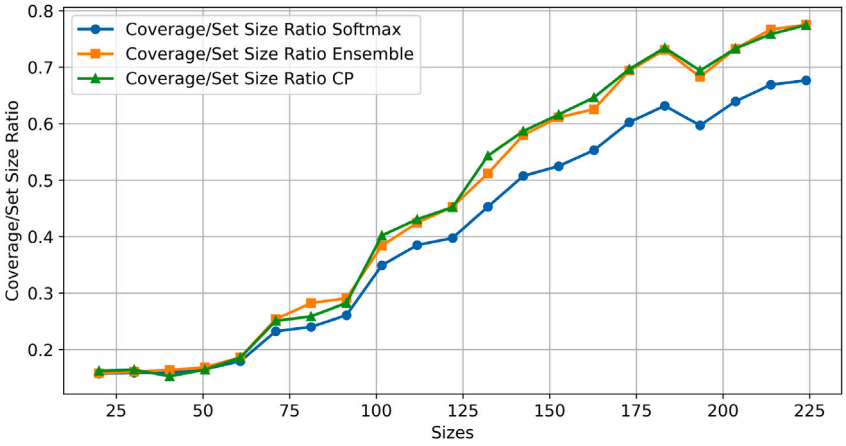


Fig. 13. Coverage/prediction set size ratio during covariate shift. ICP demonstrates superior performance compared to ensembles and softmax in the low shift region.

3. **Uncertainty decomposition:** Conformal prediction focuses on total predictive uncertainty without distinguishing between data and model uncertainty. This distinction is critical for tasks like detecting data defects or selecting informative samples in active learning.
   Challenges: Theoretically, while the total predictive uncertainty can be conceptually split into data and model components, Gruber et al. (2023) highlights the unavoidable linkage between the two. Hüllermeier and Waegeman (2019) emphasizes the need to redefine aleatoric and epistemic uncertainties based on the specific problem framework, as their definitions can shift with changes in the setup. Furthermore, Hüllermeier (2022), de Jong et al. (2024) demonstrate that common metrics struggle to reliably distinguish between the two types of uncertainty. Finally, Mucsányi et al. (2024) shows that most existing methods fail to achieve true separation.
   Promising solutions: Mucsányi et al. (2024) recommends task-specific method selection to address these challenges. New metrics proposed by Hofman et al. (2024b), Kotelevskii and Panov (2024) in the Bayesian framework show promise for better disentanglement. Integrating conformal prediction with these approaches could enable uncertainty decomposition within the conformal framework.

7. General practical implications for agricultural decision support

Inductive Conformal Prediction has significant potential in enhancing decision-making for critical agricultural tasks. By providing calibrated uncertainty estimates, ICP ensures robust error control across diverse scenarios, such as optimizing planting schedules and managing nutrient application. This reliability enables machine learning models to deliver actionable insights with clear confidence indicators, increasing stakeholder trust and reducing the risk of poor decisions.

One of ICP's major advantages is its distribution-free nature, meaning it does not rely on assumptions about the underlying data distribution. In real-world agricultural applications, data often follows complex empirical distributions that cannot be adequately captured by predefined models. ICP therefore stands in contrast to Bayesian methods, which rely on prior assumptions and likelihood functions—if the likelihood is misspecified, the posterior distribution may not reflect reality.

Additionally, ICP is computationally efficient and model-agnostic, making it accessible to users with limited computing resources, including farmers. ICP enables reliable model democratization, allowing local farmers to make informed decisions without expensive infrastructure. Another key advantage is calibrated prediction intervals using limited data. ICP reduces the need for large datasets while maintaining reliable uncertainty estimates. In cases where recalibration is required due to distributional shifts – such as daily weather variations or seasonal changes – ICP can achieve predefined error levels with as few as 1000 labeled samples. This significantly lowers labeling costs. Finally, ICP provides transparent and interpretable uncertainty representation, which is particularly beneficial for users without deep statistical expertise. For example, a prediction interval such as $Harvest$ in5 ± 2 days is more intuitive than probability scores, making decision-making more straightforward for agricultural practitioners.

Table 8
Qualitative comparison of uncertainty quantification techniques.

Method	Model Agnostic	Calibrated	OoD Detection	Distributional Shift	Computational Complexity	Inference Time	Measured Uncertainty
Conformal prediction	Yes	Yes	Limited	Limited	Low	low	TU
Bayesian	No	No	Yes	Yes	High	High	TU, AU, & EU
Softmax	limited	No	limited	No	Low	Low	AU

TU: Total predictive uncertainty, **AU:** Aleatoric uncertainty, and **EU:** Epistemic uncertainty.

8. Conclusion

Our study demonstrates the potential of Inductive Conformal Prediction (ICP) to significantly enhance uncertainty quantification (UQ) in machine learning models, particularly within agricultural decision support systems. By converting point predictions into uncertainty-aware outputs, ICP facilitates more reliable and informed decision-making, which is crucial in high-stakes environments.

The model-agnostic, computationally efficient, and distribution-free nature of ICP, along with its ability to provide calibrated uncertainty estimates, positions it as a valuable tool for improving the safety and effectiveness of machine learning applications across diverse domains. Our ablation study emphasizes the need for a balanced approach when tuning ICP parameters, particularly in managing the trade-off between precision and conservatism.

The case study on the sugar beet dataset illustrates how ICP can diagnose issues in data, models, or both, showcasing the importance of domain knowledge in setting the confidence level (α). While ICP shows acceptable performance under covariate shift, its limitations in out-of-distribution (OoD) detection suggest that careful selection of UQ methods is necessary based on specific application contexts.

Despite these limitations, our research highlights promising future directions, including enhancing OoD detection, exploring uncertainty decomposition, and refining the calibration dataset requirements. The impact of conformal prediction on digital agriculture is particularly notable, offering a straightforward, calibrated UQ model that can be effectively utilized by agricultural practitioners. By promoting more transparent and reliable AI-driven systems, our approach contributes to advancing digital agriculture toward greater efficiency and sustainability.

CRediT authorship contribution statement

Mohamed Farag: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Ahmed Emam:** Methodology, Investigation, Writing – review & editing. **Johannes Leonhardt:** Methodology, Investigation, Writing – review & editing. **Ribana Roscher:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 - 390732324 - PhenoRob. Further, it has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - RO 4839/6-1 - 459376902 - AID4Crops, RO 4839/5-1 / SCHM 3322/4-1 - 458156377 - MapInWild, and SFB 1502/1-2022 - 450058266 - DETECT.

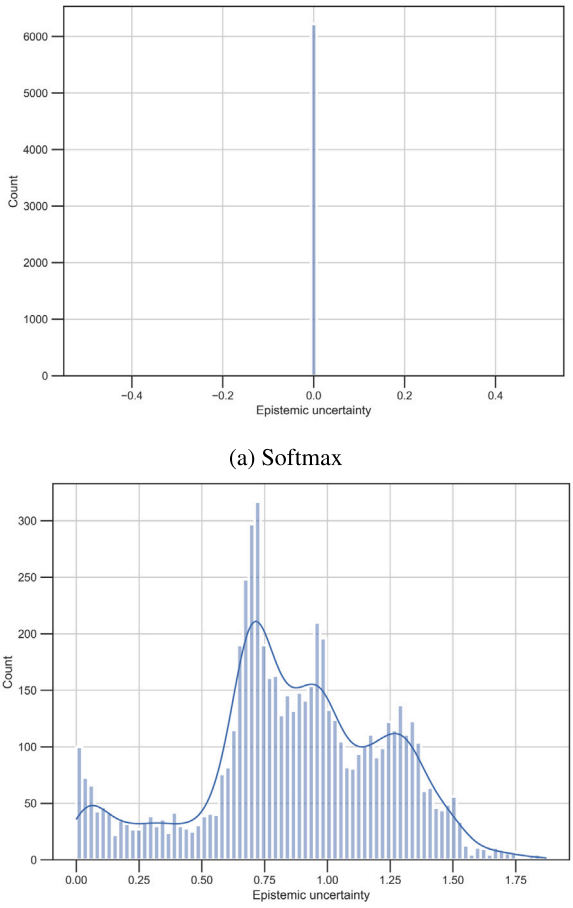


Fig. 14. Epistemic uncertainty representation for GrowliFlower as OoD. The single model exhibits overconfidence for 100% of the data, as indicated by the zero bin which highlights the limitations of using a single model (softmax outputs) for OoD detection. In contrast, the five members ensemble exhibits overconfidence for only 1.6% of the data represented by the most left pin, a result of the coarse approximation to the true model which could be enhanced by increasing the number of ensemble members.

Appendix

Table A.9 shows the performance of different models used in this study. We do not observe significant performance when we use single model against ensembles on the predictive performance. However, the difference in the uncertainty estimates and the calibration. Fig. 14 demonstrates the ability of the ensembles to use the epistemic uncertainty to show more uncertainty for OoD samples compared to the softmax approach.

For OoD detection experiment, At normal conditions where we have testing set from same distribution as DND-SB, we achieve average confidence around 98% and credibility 44.45%. Hence, we trust the predictions. While for GrowliFlower as OoD, we get 38.8% average credibility and 97.7% confidence. The reduction in credibility by 12.7%

Table A.9
Performance metrics. The upper section shows results for the DND-SB dataset, and the lower section shows results for GrowliFlower.

Dataset/Metric	ViT-B/16	ResNet-18	ResNet-18 ensemble ($M = 5$)	ViT-B/16 ensemble ($M = 5$)
DND-SB				
Acc_{test}	91.8%	92.2%	92.2%	86.9%
GrowliFlower				
Acc_{test}	74.7%	71.1%	69.6%	70.6%

is another sign that GrowliFlower samples are different compared to DND-SB calibration samples. However, we aim to get credibility no more than 5.0% for different datasets to reliably identify it as OoD.

We attempt to adjust by setting α at values of 0.005, 0.05, 0.5, and 0.8, increasing the precision instead of controlling miscoverage because, in this context, miscoverage loses its relevance when dealing with an entirely new dataset/task observing an increase in null sets. However, at the $\alpha = 0.8$ level, we are unable to completely eliminate the singleton sets.

Data availability

I have shared the links for datasets, code will be available on request.

References

Agyeman, P.C., Kebonye, N.M., Khosravi, V., Kingsley, J., Borůvka, L., Vašát, R., Boateng, C.M., 2023. Optimal zinc level and uncertainty quantification in agricultural soils via visible near-infrared reflectance and soil chemical properties. J. Environ. Manag. 326, 116701. http://dx.doi.org/10.1016/j.jenvman.2022.116701.

van Amersfoort, J.R., Smith, L., Teh, Y.W., Gal, Y., 2020. Uncertainty estimation using a single deep deterministic neural network. In: International Conference on Machine Learning.

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

Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T., 2024. Conformal risk control. In: The Twelfth International Conference on Learning Representations.

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2022. Conformal prediction beyond exchangeability. Ann. Statist. URL: https://api.semanticscholar.org/CorpusID:247158820.

Cabezas, L.M.C., Santos, V.S., Ramos, T., Izbicki, R., 2025. Epistemic Uncertainty in Conformal Scores: A Unified Approach. URL: https://api.semanticscholar.org/CorpusID:276258945.

Celikkan, E., Saberioon, M., Herold, M., Klein, N., 2023. Semantic segmentation of crops and weeds with probabilistic modeling and uncertainty quantification. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops. ICCVW, pp. 582–592. http://dx.doi.org/10.1109/ICCVW60793.2023.00065.

Chrispell, J.C., Jenkins, E.W., Kavanagh, K.R., Parno, M.D., 2021. Characterizing prediction uncertainty in agricultural modeling via a coupled statistical–physical framework. Modelling 2 (4), 753–775. http://dx.doi.org/10.3390/modelling2040040, URL: https://www.mdpi.com/2673-3951/2/4/40.

de Jong, I.P., Sburlea, A.I., Valdenegro-Toro, M., 2024. How disentangled are your classification uncertainties?. arXiv arXiv:2408.12175, URL: https://api.semanticscholar.org/CorpusID:271923811.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Drees, L., Demie, D.T., Paul, M.R., Leonhardt, J., Seidel, S.J., Döring, T.F., Roscher, R., 2024. Data-driven crop growth simulation on time-varying generated images using multi-conditional generative adversarial networks. Plant Methods 20 (1), http://dx.doi.org/10.1186/s13007-024-01205-3.

Drees, L., Weber, I., Rußwurm, M., Roscher, R., 2022. Time dependent image generation of plants from incomplete sequences with CNN-transformer. In: DAGM German Conference on Pattern Recognition. Springer, pp. 495–510.

FAO, IFAD, UNICEF, WFP, WHO, 2022. The State of Food Security and Nutrition in the World 2022. FAO ; IFAD ; UNICEF ; WFP ; WHO; ISSN: 2663-8061.

Farag, M., Kierdorf, J., Roscher, R., 2023. Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops. ICCVW, pp. 651–659. http://dx.doi.org/10.1109/ICCVW60793.2023.00072.

Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 145, 311–318. http://dx.doi.org/10.1016/j.compag.2018.01.009.

Food and Agriculture Organization (FAO), 2015. Status of the World's Soil Resources. Technical Report, FAO.

Gal, Y., Ghahramani, Z., 2015. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning.

Gammerman, A., Vovk, V., 2007. Hedging predictions in machine learning. Comput. J. 50 (2), 151–163. http://dx.doi.org/10.1093/comjnl/bxl065.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56 (Suppl 1), 1513–1589. http://dx.doi.org/10.1007/s10462-023-10562-9.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2), 243–268. http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x, arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437, URL: https://doi.org/10.1198/016214506000001437.

Gruber, C., Schenk, P.O., Schierholz, M., Kreuter, F., Kauermann, G., 2023. Sources of uncertainty in machine learning - A statisticians' view. arXiv arXiv:2305.16703.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017a. On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 1321–1330.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017b. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, JMLR.org, pp. 1321–1330.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hernández, S., López, J.L., 2020. Uncertainty quantification for plant disease detection using Bayesian deep learning. Appl. Soft Comput. 96, 106597. http://dx.doi.org/10.1016/j.asoc.2020.106597, URL: https://www.sciencedirect.com/science/article/pii/S1568494620305354.

Hoffmann, L., Elster, C., 2021. Deep ensembles from a Bayesian perspective. CoRR arXiv:2105.13283.

Hofman, P., Sale, Y., Hüllermeier, E., 2024a. Quantifying aleatoric and epistemic uncertainty: A credal approach. In: ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling. URL: https://openreview.net/forum?id=MhLnSoWp3p.

Hofman, P., Sale, Y., Hüllermeier, E., 2024b. Quantifying aleatoric and epistemic uncertainty with proper scoring rules. arXiv arXiv:2404.12215, URL: https://api.semanticscholar.org/CorpusID:269214579.

Hüllermeier, E., 2022. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Conference on Uncertainty in Artificial Intelligence.

Hüllermeier, E., Waegeman, W., 2019. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110, 457–506.

Kakhani, N., Alamdar, S., Kebonye, N.M., Amani, M., Scholten, T., 2024. Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. Remote. Sens. 16 (3), http://dx.doi.org/10.3390/rs16030438.

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

Kierdorf, J., Junker-Frohn, L.V., Delaney, M., Olave, M.D., Burkart, A., Jaenicke, H., Muller, O., Rascher, U., Roscher, R., 2023. GrowliFlower: An image time-series dataset for growth analysis of cauliflower. J. Field Robot. 40 (2), 173–192. http://dx.doi.org/10.1002/rob.22122.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Kirchhof, M., Kasneci, E., Oh, S.J., 2023. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In: Proceedings of the 40th International Conference on Machine Learning. ICML '23, JMLR.org.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Kotelevskii, N., Panov, M., 2024. Predictive uncertainty quantification via risk decompositions for strictly proper scoring rules. arXiv arXiv:2402.10727, URL: https://api.semanticscholar.org/CorpusID:267740331.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90. http://dx.doi.org/10.1145/3065386.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V.I., Bertin, P., Rector-Brooks, J., Korablyov, M., Bengio, Y., 2023. DEUP: Direct epistemic uncertainty prediction. Trans. Mach. Learn. Res. Expert Certification.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 6405–6416.

Lei, J., Wasserman, L., 2014. Distribution-free prediction bands for non-parametric regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 71–96.

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., Okumura, M., 2024. A survey on deep active learning: Recent advances and new frontiers. IEEE Trans. Neural Networks Learn. Syst. URL: https://api.semanticscholar.org/CorpusID:269484121.

Li, Y., Yan, S., Gong, J., 2023. Quantifying uncertainty in soil moisture retrieval using a Bayesian neural network framework. Comput. Electron. Agric. 215, 108414. http://dx.doi.org/10.1016/j.compag.2023.108414, URL: https://www.sciencedirect.com/science/article/pii/S0168169923008025.

Ling, C.X., Huang, J., Zhang, H., 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In: International Joint Conference on Artificial Intelligence. URL: https://api.semanticscholar.org/CorpusID:118673880.

Liu, J.Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Vancouver, Canada.

Lowe, N., 2021. The global challenge of hidden hunger: perspectives from the field. Proc. Nutr. Soc. 80 (3), 283–289. http://dx.doi.org/10.1017/S0029665121000902, PMID: 33896431.

Meenken, E.D., Triggs, C.M., Brown, H.E., Sinton, S., Bryant, J.R., Noble, A.D., Espig, M., Sharifi, M., Wheeler, D.M., 2021. Bayesian hybrid analytics for uncertainty analysis and real-time crop management. Agron. J. 113 (3), 2491–2505. http://dx.doi.org/10.1002/agj2.20659, URL: https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/agj2.20659, arXiv:https://acsess.onlinelibrary.wiley.com/doi/pdf/10.1002/agj2.20659.

Melki, P., Bombrun, L., Diallo, B., Dias, J., Costa, J.-P.D., 2023. Group-conditional conformal prediction via quantile regression calibration for crop and weed classification. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops. ICCVW, pp. 614–623. http://dx.doi.org/10.1109/ICCVW60793.2023.00068.

Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F., 2012. A unifying view on dataset shift in classification. Pattern Recognit. 45 (1), 521–530. http://dx.doi.org/10.1016/j.patcog.2011.06.019.

Mucsányi, B., Kirchhof, M., Oh, S.J., 2024. Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks. Adv. Neural Inform. Process. Syst. 37, 50972–51038.

Mukhoti, J., Gal, Y., 2018. Evaluating Bayesian deep learning methods for semantic segmentation. arXiv arXiv:1811.12709.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H., Gal, Y., 2023. Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 24384–24394.

Mustafa, G., Hayat, N., Alotaibi, B.A., 2023. How and why to prevent over fertilization to get sustainable crop production. Sustain. Plant Nutr. 339–354. http://dx.doi.org/10.1016/B978-0-443-18675-2.00019-5, URL: https://www.sciencedirect.com/science/article/pii/B9780443186752000195.

Nalisnick, E., Matsukawa, A., Teh, Y.W., Rezende, D.J., Blei, D.M., 2019. Do deep generative models know what they don't know? In: Advances in Neural Information Processing Systems (NeurIPS).

Neupane, B., Horanont, T., Hung, N.D., 2019. Deep learning based banana plant detection and counting using high-resolution red-green-blue (RGB) images collected from unmanned aerial vehicle (UAV). PLoS ONE 14.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. Ser. A, Contain. Pap. A Math. Or Phys. Character 231 (694–706), 289–337.

Nix, D.A., Weigend, A.S., 1994. Estimating the mean and variance of the target probability distribution in neural network training. In: Proceedings of the IEEE International Conference on Neural Networks. ICNN, pp. 55–60.

Novello, P., Dalmau, J., Andeol, L., 2024. Out-of-distribution detection should use conformal prediction (and vice-versa?). arXiv arXiv:2403.11532, URL: https://api.semanticscholar.org/CorpusID:268531401.

Ovadia, Y., Fertig, E., Ren, J.J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Neural Information Processing Systems.

Padarian, J., Minasny, B., McBratney, A., 2022. Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout. Geoderma 425, 116063. http://dx.doi.org/10.1016/j.geoderma.2022.116063, URL: https://www.sciencedirect.com/science/article/pii/S0016706122003706.

Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., 2002. Inductive confidence machines for regression. In: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), Machine Learning: ECML 2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 345–356.

Papamarkou, T., Skoularidou, M., Palla, K., Aitchison, L., Arbel, J., Dunson, D.B., Filippone, M., Fortuin, V., Hennig, P., Hern'andez-Lobato, J.M., Hubin, A., Immer, A., Karaletsos, T., Khan, M.E., Kristiadi, A., Li, Y., Mandt, S., Nemeth, C., Osborne, M.A., Rudner, T.G.J., Rugamer, D., Teh, Y.W., Welling, M., Wilson, A.G., Zhang, R., 2024. Position: Bayesian deep learning is needed in the age of large-scale AI. In: International Conference on Machine Learning. URL: https://api.semanticscholar.org/CorpusID:270094572.

Postels, J., Segu, M., Sun, T., Gool, L.V., Yu, F., Tombari, F., 2021. On the practicality of deterministic epistemic uncertainty. In: International Conference on Machine Learning.

Pound, M.P., Atkinson, J.A., Townsend, A.J., Wilson, M.H., Griffiths, M., Jackson, A.S., Bulat, A., Tzimiropoulos, G., Wells, D.M., Murchie, E.H., et al., 2017. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience 6 (10), gix083.

Quan, L., Feng, H., Lv, Y., Wang, Q., Zhang, C., Liu, J., Yuan, Z., 2019. Maize seedling detection under different growth stages and complex field environments based on an improved faster R–CNN. Biosyst. Eng. 184, 1–23. http://dx.doi.org/10.1016/j.biosystemseng.2019.05.002.

Rau, K., Eggensperger, K., Schneider, F., Hennig, P., Scholten, T., 2024. How can we quantify, explain, and apply the uncertainty of complex soil maps predicted with neural networks? Sci. Total Environ. 944, 173720. http://dx.doi.org/10.1016/j.scitotenv.2024.173720.

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., Verri, A., 2004. Are loss functions all the same? Neural Comput. 16 (5), 1063–1076. http://dx.doi.org/10.1162/089976604773135104.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., Sabokrou, M., 2021. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. Trans. Mach. Learn. Res. 2022.

Shaker, M.H., Hüllermeier, E., 2021. Ensemble-based uncertainty quantification: Bayesian versus credal inference. In: Proceedings 31. Workshop Computational Intelligence. http://dx.doi.org/10.48550/ARXIV.2107.10384.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

Stutz, D., Roy, A.G., Matejovicova, T., Strachan, P., Cemgil, A.T., Doucet, A., 2024. Conformal prediction under ambiguous ground truth. Transactions on Machine Learning Research.

Teixeira, I., Morais, R., Sousa, J.J., Cunha, A., 2023. Deep learning models for the classification of crops in aerial imagery: A review. Agriculture 13 (5), http://dx.doi.org/10.3390/agriculture13050965.

Verdoja, F., Kyrki, V., 2020. Notes on the behavior of MC dropout. http://dx.doi.org/10.48550/arXiv.2008.02627.

Vladimir Vovk, G.S., 2005. Algorithmic Learning in a Random World, first ed. Springer New York, NY, http://dx.doi.org/10.1007/b106715.

Vovk, V., 2012a. Conditional validity of inductive conformal predictors. In: Hoi, S.C.H., Buntine, W. (Eds.), Proceedings of the Asian Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 25, PMLR, Singapore Management University, Singapore, pp. 475–490.

Vovk, V., 2012b. Cross-conformal predictors. Ann. Math. Artif. Intell. 74, 9–28, URL: https://api.semanticscholar.org/CorpusID:51973287.

Vovk, V., Gammerman, A., Shafer, G., 2022. Algorithmic learning in a random world, second ed. In: Mathematics and Statistics, Springer, Switzerland, http://dx.doi.org/10.1007/978-3-031-06649-8, Hardcover ISBN: 978-3-031-06648-1. Published: 14 December 2022. Softcover ISBN: 978-3-031-06651-1. Due: 28 December 2023. eBook ISBN: 978-3-031-06649-8. Published: 13 December 2022..

Wang, H., Wellmann, F., Zhang, T., Schaaf, A., Kanig, R.M., Verweij, E., von Hebel, C., van der Kruk, J., 2019. Pattern extraction of topsoil and subsoil heterogeneity and soil-crop interaction using unsupervised Bayesian machine learning: An application to satellite-derived NDVI time series and electromagnetic induction measurements. J. Geophys. Res.: Biogeosciences 124 (6), 1524–1544. http://dx.doi.org/10.1029/2019JG005046, URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JG005046, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019JG005046.

Wang, H., Yeung, D.-Y., 2020. A survey on Bayesian deep learning. ACM Comput. Surv. 53 (5), http://dx.doi.org/10.1145/3409383.

Welling, M., Teh, Y.W., 2011. Bayesian learning via stochastic gradient langevin dynamics. In: International Conference on Machine Learning. URL: https://api.semanticscholar.org/CorpusID:2178983.

Wen, Y., Vicol, P., Ba, J., Tran, D., Grosse, R.B., 2018. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In: International Conference on Learning Representations. ICLR, arXiv:1803.04386.

Werther, M., Odermatt, D., Simis, S.G., Gurlin, D., Lehmann, M.K., Kutser, T., Gupana, R., Varley, A., Hunter, P.D., Tyler, A.N., Spyrakos, E., 2022. A Bayesian approach for remote sensing of chlorophyll-a and associated retrieval uncertainty in oligotrophic and mesotrophic lakes. Remote Sens. Environ. 283, 113295. http://dx.doi.org/10.1016/j.rse.2022.113295, URL: https://www.sciencedirect.com/science/article/pii/S0034425722004011.

Xu, L., Chen, N., Yang, C., Yu, H., Chen, Z., 2022. Quantifying the uncertainty of precipitation forecasting using probabilistic deep learning. Hydrol. Earth Syst. Sci. 26 (11), 2923–2938. http://dx.doi.org/10.5194/hess-26-2923-2022, URL: https://hess.copernicus.org/articles/26/2923/2022/.

Yi, J., Krusenbaum, L., Unger, P., Hüging, H., Seidel, S.J., Schaaf, G., Gall, J., 2020. Deep learning for non-invasive diagnosis of nutrient deficiencies in sugar beet using RGB images. Sensors 20 (20), http://dx.doi.org/10.3390/s20205893.

Zabawa, L., Kicherer, A., Klingbeil, L., Töpfer, R., Kuhlmann, H., Roscher, R., 2020. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 164, 73–83. http://dx.doi.org/10.1016/j.isprsjprs.2020.04.002.