

Assessing Conformal Prediction for Remote Sensing: Application to Local Climate Zone Classification

Christoph Schweden[✉], Peter Jung[✉], and Xiao Xiang Zhu[✉], *Fellow, IEEE*

Abstract—While neural networks are at the forefront of solving a variety of tasks in remote sensing, these black-box models often lack a clear sense of uncertainty. Currently experiencing a rebirth in the light of deep learning is the statistically founded uncertainty framework of conformal prediction (CP). Under certain weak assumptions, this framework provides theoretical coverage guarantees for the ground-truth value or class. The coverage guarantee holds for the predictions of the CP framework, which take the form of a prediction set containing one or more classes for classification or segmentation problems, or a prediction interval in the case of regression problems. Additionally, the size of the prediction set or interval also serves as an intuitive indicator of uncertainty. We investigate this uncertainty framework for a large-scale land cover classification dataset termed *So2Sat LCZ42*, which comes with ten independent expert annotations per annotated satellite image patch. We limit ourselves to the building classes of the dataset, which yield the largest proportion of expert uncertainty. The multiple annotations are here encoded into a soft label used for label uncertainty-aware training. First, we evaluate different adaptations of CP calibration methods. Next to the analysis of coverage and prediction set sizes, as a novelty, we also evaluate the methods with respect to the explainability of the conformalizer’s uncertainty toward aleatoric and epistemic uncertainty. Additionally, we construct a novel active learning (AL) framework based on CP prediction sets and show promising results in terms of expressiveness regarding predictive uncertainty. CP-based AL achieves better test performance than random querying, and yields lower predictive uncertainties than the model trained on the full training set while upholding balanced test accuracy, with only a fraction of the data needed.

Index Terms—Calibration, conformal prediction (CP), deep learning, land cover classification, uncertainty quantification (UQ).

Received 30 May 2025; revised 3 April 2026 and 3 July 2026; accepted 28 July 2026. Date of publication 4 August 2026; date of current version 18 August 2026. This work was supported in part by German Federal Ministry of Education and Research (BMBF) in the Framework of the International Future AI Laboratory “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001; in part by the Helmholtz Association through the Joint Research School Munich School for Data Science—MUDS; in part by the German Federal Ministry for Economic Affairs and Climate Action in the Framework of the “National Center of Excellence ML4Earth” under Grant 50EE2201C; and in part by Munich Center for Machine Learning. (Corresponding author: Xiao Xiang Zhu.)

Christoph Schweden is with the Department of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany.

Peter Jung is with the Institute of Optical Sensor Systems, German Aerospace Center, 82234 Berlin, Germany, and also with the Communications and Information Theory Group, Technical University of Berlin, 10623 Berlin, Germany.

Xiao Xiang Zhu is with the Department of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with Munich Center for Machine Learning, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2026.3719780

NOMENCLATURE

$\mathbf{x}$	Input data.
X	Input space.
y	Ground-truth label.
$\mathcal{Y} = \{1, \dots, K\}$	Set of class labels.
$\mathbf{Y}_i$	Vote vector for sample i .
Y_{ik}	Number of votes for class k on sample i .
V_{ij}	Vote of expert j for sample i .
V_i	Set of voted classes for sample i .
$\mathbf{y}_{\text{distr},i}$	Empirical label distribution (vector).
$\mathbf{y}_{\text{maj}}$	Majority-voted label.
M	Number of annotators.
K	Number of classes.
n	Total number of samples.
n_{calib}	Size of calibration set.
n_{test}	Size of test set.
$f_{\theta}(\mathbf{x})$	Neural network with parameters θ .
θ	Model parameters.
$\hat{\pi}_k(\mathbf{x})$	Predicted probability for class k .
$\hat{p}$	Model confidence.
$\hat{y}$	Predicted class label.
$C(\mathbf{x})$	Conformal prediction set for input $\mathbf{x}$.
α	Miscoverage level.
$\hat{t}$	Conformal threshold (quantile).
$s(\mathbf{x}, y)$	Conformal score function.
$\hat{\pi}_{\text{ord},j}(\mathbf{x})$	j th largest ordered probability.
λ	Regularization parameter (RAPS).
k_{reg}	Regularization cutoff parameter (RAPS).
$\sigma_x(\mathbf{y})$	Rank of label y in ordered probabilities.
Cov	Coverage of prediction sets.
Ineff	Inefficiency (average prediction set size).
$ C(\mathbf{x}) $	Size of prediction set.
$\text{Cov}_{\text{lab},i}$	Label coverage for sample i .
$\text{IoU}_{\text{pred-vot},i}$	IoU between prediction set and votings.
$\text{Cov}_{\text{lab-vot},i}$	Vote-weighted label coverage.
$\text{IoU}_{\text{lab-vot},i}$	Vote-weighted IoU.
$H(\mathbf{y}_{\text{distr},i})$	Shannon entropy (label uncertainty).

I. INTRODUCTION

UNCERTAINTIES can arise from many different sources. In remote sensing, these can, for example, be various types of sensor noise [1] during data curation or uncertainties in the data processing pipeline. When remote sensing data is then labeled in the next step, this assignment can also contain uncertainty [2]. We can think of measurement uncertainty regarding the target value or ambiguity among labels for a classification setting. Generally speaking, all of these

components form the aleatoric or data uncertainty. This type of uncertainty is inherent in the data and cannot be modeled away. The epistemic or model uncertainty, on the other hand, arises from model misspecification or lack of adaptation. Since the era of deep learning, many frameworks for quantifying the two sources of uncertainty have been established [3]. While there are both frequentist and Bayesian approaches, which both yield competitive performance, ensemble techniques [4] have also been established that combine the predictions of multiple ensemble members. For classification problems, there is a straightforward experiment to evaluate the model predictions: Comparing the model's confidence (typically the highest softmax pseudo-probability) to its accuracy binwise is usually referred to as a calibration plot or reliability diagram. The analysis can give a first indication about the quality of the predictive probabilities and hence also predictive uncertainties based on them. The confidence is also referred to as maximum softmax probability (MSP), and 1 minus the MSP often is utilized as a measure of uncertainty in out-of-distribution detection scenarios [5], [6].

A novel research direction in the context of uncertainty quantification (UQ) has been opened by [7], who introduced *conformal prediction* (CP). Broadly speaking, any predictor can be modified using CP such that the true value or class is predicted with a prespecified confidence. In more detail, this means that the modified predictor does not yield a single point prediction anymore but instead predicts a set of classes in the case of classification or a prediction interval in the case of regression. A core distinction can be made between transductive and inductive CP: While transductive [8] approaches require to evaluate the nonconformity measure on all possible subsets of labels within the training set, inductive [9] methods solve this issue by requiring a separate hold-out calibration set on which the measure is evaluated. Here, it is essential that the calibration set is *exchangeable* [10] solely with the testing set [11]. This is equivalent to the fact that any finite permutation of a subset coming from one of the three sets (or a mix thereof) follows the same joint probability distribution as the original subset. Note that this assumption is similar to the *i.i.d.* (independent, identically distributed) assumption of deep learning frameworks, and one can link the two assumptions using De Finetti's theorem [12]. Furthermore, the conformalized model also inherits a notion of predictive uncertainty next to the theoretically proven coverage guarantee of the true class or point. This notion is typically stated by the size of the prediction set in the classification case or, for regression problems, by the width of the prediction interval.

For the evaluation of predictive uncertainties, one usually has to construct a downstream experiment that explicitly uses the quantified uncertainties. This is required because different standalone uncertainty estimates are rarely comparable in a meaningful way. One such downstream task is active learning (AL), where a model is initially trained with a relatively small subset of the existing training dataset. Once trained, the model's predictive uncertainties on the remainder of the training dataset are utilized to sequentially add data to the initial dataset pool. In an early work, Angluin [14] describes this process of *querying* data, which is the key concept in AL. Querying here means nothing other than handing certain

identified data points to a so-called human oracle, which then reveals the ground-truth label or target value. The model is then retrained on these newly added labeled data points. The overall goal of doing this iteratively is to quickly adapt the model by feeding it with a few representative data samples at a time. Here, the quality of the predictive uncertainties is essential as it determines the representativeness of the predicted data points. What remains crucial is to evaluate the uncertainty-based querying strategies against the random baseline, i.e., adding random points to the training dataset pool. Overall, typically the accuracy on a hold-out testing dataset is evaluated after each retraining step of the model. In the proposed experimental setting, predictive uncertainty is further considered as an objective variable.

The dataset being evaluated in this work follows the local climate zone (LCZ) class scheme [15] shown in Fig. 1. Initially proposed for studying urban heat islands, the scheme has since been integrated into many urban climate applications such as green space planning [16] or population estimation [17]. Though many geographic information system (GIS)-based mapping studies have been presented (see [18] for a recent survey), the focus has shifted to deep learning-based methods recently. Exemplarily, residual convolutional neural networks with multiseasonal data [19], residual and nonlocal networks [13], and object-based convolutional neural networks [20] have been used. To summarize, we make the following main contributions.

- 1) We evaluate the promising CP framework by means of various adapted score functions on a large-scale land cover benchmark dataset.
- 2) We investigate the prediction sets with respect to the coverage as well as their explainable power toward aleatoric and epistemic uncertainty.
- 3) We build a novel CP score-based AL benchmark and show promising results that beat the random baseline by a large margin, both with respect to accuracy and predictive uncertainty.

In the following work, we proceed as follows. First, Section II gives an overview of current and past research in the fields of UQ, CP, and AL, as well as applications of UQ to the field of remote sensing. Then, Section III introduces the dataset and the neural network utilized, before the general CP framework for classification is introduced. Furthermore, the means of evaluating CP methods are introduced, and an AL framework based on CP is explained. Next, Section IV shows experimental results on the *So2Sat LCZ42* dataset, namely a benchmark of different CP methods with respect to coverage guarantees, prediction set size, and expert label votes coverage, as well as an analysis at the class level. Lastly, Section V presents the AL experimental results based on a CP-based querying function, before Section VI concludes. The code can be found here: https://github.com/ChrisKo94/CP_LCZ

II. RELATED WORK

A variety of streams have evolved to tackle UQ in deep learning; see [21] for a recent overview. Although creating a significant computational overhead, ensembles of networks

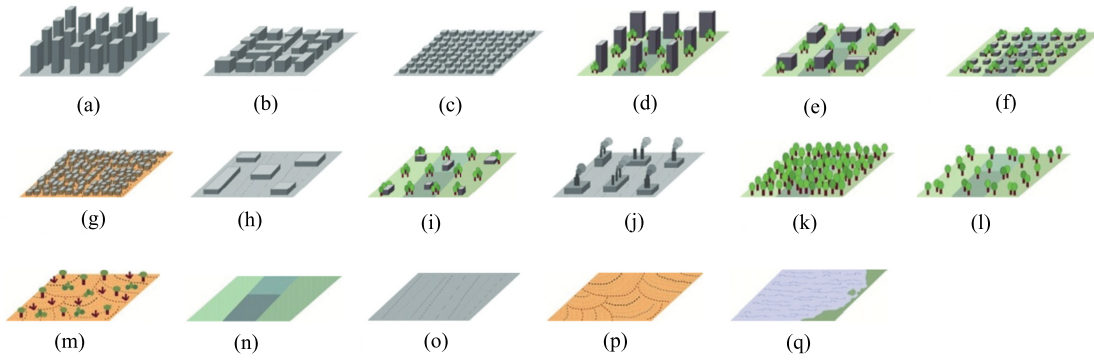


Fig. 1. LCZ classification scheme as presented by [13]. (a) Compact high-rise. (b) Compact mid-rise. (c) Compact low-rise. (d) Open high-rise. (e) Open mid-rise. (f) Open low-rise. (g) Lightweight low-rise. (h) Large low-rise. (i) Sparsely built. (j) Heavy industry. (k) Dense trees. (l) Scattered trees. (m) Bush, scrub. (n) Low plants. (o) Bare rock or paved. (p) Bare soil or sand. (q) Water.

[4] have been established as a hard-to-beat baseline in many applications. Using a Bayesian setting, on the other hand, allows for relaxing the solution space of a network by placing prior distributions on the weights [22], [23]. A prominent Bayesian approximation technique is stated by *MC Dropout* [24], which uses multiple forward passes through a network with dropout to yield predictive uncertainties. Deterministic networks have been studied as another stream of UQ methods. Here, the uncertainties are based on a single forward pass through a deterministic network. Popular approaches are *SNGP* [6], which combines spectral normalization of weight matrices and an approximate Gaussian process (GP) as output layer, and *DUE* [25], which introduces spectral batch normalization and enriches the aforementioned GP with a more stable inducing point approximation. As a substream of the deterministic approaches, Dirichlet-based approaches have been introduced. These utilize Dirichlet distributions to loosen the typical categorical output space in classification problems. Prominent representatives are the *Prior Networks* [26] and the *Evidential Networks* [27].

Since its introduction by [7], CP has been developed methodologically and applied to many research fields. In particular for deep learning models, many adaptations have been introduced recently [28]. A first milestone approach for classification tasks is given by the *least ambiguous set-valued classifier (LAC)* [29]. Here, the foundation was laid for constructing prediction sets that satisfy the theoretical coverage guarantees and, at the same time, are least ambiguous (i.e., as small as possible). As an advancement of LAC, *adaptive prediction sets (APSs)* were introduced soon after by [30]. Via generalizing the quantile function of the LAC approach and adding a subtraining routine to the calibration step, the authors showed increased conditional coverage across various datasets.

By adding a regularization mechanism, [31] constructed *regularized APSs (RAPs)* and were able to further reduce the prediction set sizes while still fulfilling the theoretical coverage guarantees. The authors further introduced a more straightforward prediction set constructor, namely the *Top-k* algorithm, which simply considers the k most likely predictions where k is again estimated in such a way that it fulfills the prespecified coverage guarantee. Recent methodological

advances include, exemplarily, the adaptation of CP to cases beyond the exchangeability assumption [32] or linking CP with an early stopping mechanism [33]. Works on fusing CP with Monte Carlo Dropout [34] or with temperature scaling [35] were also presented recently.

AL was introduced to machine learning algorithms by [36], who adopted the framework to support vector machines (SVMs). Since then, many works, particularly involving deep learning architectures, have been published; see [37] for a recent survey. In the field of remote sensing, AL also plays a pivotal role. An early application using SVMs was presented by [38], who used AL for the task of segmenting multispectral images. Further works focused on hyperspectral image classification [39], remote sensing image classification in a more general sense involving multiple algorithms and datasets [40], batch-mode AL [41], and recently also on a multilabel classification framework [42].

AL in combination with CP has also been investigated. Ho and Wechsler [43] initially proposed a transductive CP framework for querying, and Balasubramanian et al. [44] added an eigendecomposition for the indication of information to the framework. An inductive approach, on the other hand, was proposed by [45] for an ensemble of neural networks. Matiz and Barner [46] constructed a testbed for image classification using notions of informativeness, diversity, and information density in a single query function. The same authors later also proposed another query function based on a constrained linear regression model, which further incorporates a notion of uncertainty [47].

In the field of remote sensing, uncertainties continue to play an important role. Systematic quantification and evaluation, however, are often not in focus. The uncertainty evaluation downstream task of out-of-distribution detection was studied both by [48], who utilized modified versions of prior networks, and by [49], where distance-aware neural networks were considered. Calibration can be seen as another way of evaluating uncertainties in deep learning [50]. Koller et al. [51] investigated the calibration and generalization performance of a remote sensing dataset, which has been independently labeled by ten experts, a continuation of the research stream on human label uncertainty [52]. The expert labels of this dataset were also studied by [2], who utilized a multinomial

mixture model and a stochastic version of the Expectation-Maximization algorithm, assuming no label ambiguity.

Recently, CP has also found potential applications in remote sensing. Bhattacharjee and Jung [53] studied the effect of label noise on the calibration results of various CP methods, while [54] investigated three large-scale use-cases for CP in remote sensing image analysis. Regarding land use land cover (LULC) classification, recent studies investigated conformal statistics for pixel-wise uncertainty [55], uncertainty maps based on class-conditional CP [56], as well as trustworthy hyperspectral image classification [57]. Note, however, that no comparison was drawn to existing data uncertainty and no downstream applications were employed using CP, which marks the novelty of this work. Further CP applications include soil organic carbon estimation [58], birds and drone detection [59], as well as harvest readiness classification [60] and tropical cyclone track forecasting [61].

Regarding LCZ classification, recent advances have been extensively surveyed by [62]. While GIS-based LCZ classification is often rule-based, some recent works also employed Bayesian frameworks [63] or clustering algorithms [64]. Regarding remotely sensed data, recent trends in LCZ classification have evolved in the direction of multisource data [65] or optimizing the combination of model and data source [66]. However, recently, a combination of GIS and remote sensing data has been utilized [67]. As a novelty, we incorporate CP for the tasks of explainability and AL.

III. METHODOLOGY

A. Data and Model

The notation used in the following is depicted in the nomenclature. The dataset that will be studied in this work follows the classification scheme of the *LCZs* introduced by [15]. This 17-class scheme describes urban conglomerates and their surroundings and can be further categorized into urban and vegetation classes. A dataset with global coverage of 42 cities was introduced to the community by [68], who manually labeled co-registered Sentinel-1 and Sentinel-2 patches according to the LCZ labeling scheme. Note that the dataset does not cover the entire cities, but focuses on homogeneous regions which can be attributed to a single class. Nonetheless, significant ambiguity exists among the classes. This ambiguity can be attributed to the characterization of the different classes, which exemplarily depend on the density and height of the buildings in the case of the urban classes. This ambiguity motivates the use of CP, as a prediction set with coverage guarantee can help to better explain the model's uncertainty when dealing with ambiguous classes.

Next to the global dataset, a subset of the data was also investigated to determine a notion of label confidence. For this, ten remote sensing experts were asked independently to label the identical data coming from ten European cities. By doing so, a label confidence (i.e., average agreement rate among the experts) of ca. 85% was estimated. The full dataset contains ca. 240 thousand labeled patches of size 32 by 32 pixels and

can be found online.¹ We denote the individual label votes of the ten experts for image i via

$$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iK}) \text{ where } Y_{ik} = \sum_{j=1}^M \mathbb{1}_{\{V_{ij}=k\}} \quad (1)$$

and $V_{ij} \in \{1, \dots, K\}$ denotes a vote from expert j on image i for class k ($K = 17$ denotes the number of LCZ classes). We furthermore normalize the vote vectors, yielding the empirical distribution of the votes given by

$$\mathbf{y}_{\text{distr},i} = \mathbf{Y}_i / M \quad (2)$$

where $M = 10$ denotes the number of voters. The distributional (soft) label defined via the empirical distribution (2) will be used for training the models later on in order to be label uncertainty-aware. Although moderate ambiguity exists among the experts, for each image, the majority-voted class is uniquely defined. We denote this class by y_{maj} and treat it as the ground-truth class where applicable.

The 17 classes can be divided into ten building and seven nonbuilding classes, see Fig. 1. Since the agreement of experts for the nonbuilding classes was much higher on average (and those classes are generally easier to classify and yield very little ambiguity), we decided to solely target the building classes with our CP framework in order to focus on the more expressive classes in terms of human label uncertainty. Little to no ambiguity exists among building and nonbuilding classes (see [68] for details). Therefore, we manually excluded the nonbuilding samples from the dataset without having to sacrifice label information from the building classes.

We are thus left with ca. 60 thousand labeled Sentinel-2 images, which we will utilize in the following experiments. As shown by [69], the relatively small patch size of these data samples does not require a particularly deep neural network for effectively learning the labels. The authors introduced *Sen2LCZ*, a lightweight convolutional neural network that outperformed classical deep architectures such as a ResNet [70] or a VGG [71]. Due to this fact and later effective usages of this model (e.g., [48]), we choose this model. It is noteworthy, however, that ResNets can be made theoretically *distance-aware* [6], which makes them particularly suitable for Out-of-Distribution detection tasks in land cover classification [49], [53] due to the popularity of the ResNet as a classifier.

Previous advancements by [51] led to our choice of the *Kullback-Leibler (KL) Divergence* as a loss function in combination with the soft labels earlier defined. It could be shown that these changes already lead to significant improvements regarding calibration; hence, we refrained from further calibration techniques. We note, however, that further improvements in CP could be theoretically achieved with even better calibrated probabilities. Hyperparameters and model specifications, such as early stopping and learning rate decay, were chosen similar to previous works. For a more realistic testing scenario, we decided to use the train, calibration, and test split proposed by [51]: Training data comes from a different set of European cities than calibration and test data, but since the

¹So2Sat LCZ42 Evaluation dataset (Link to Download: <https://mediatum.ub.tum.de/1659039>)

latter two are coming from the same cities, exchangeability between calibration and test data is assured. Nevertheless, there exists a significant generalization gap to bridge between training and testing.

B. General CP Framework for Classification

We now introduce the general framework for inductive CP in the case of a deep neural network classifier. Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n} \in (\mathcal{X} \times \mathcal{Y})^n$ be the data consisting of data points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ and labels $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathcal{Y} = \{1, \dots, K\}$ coming from the sets $\mathcal{X}$ and $\mathcal{Y}$. We learn the mapping between labels and data points using a neural network $f_\theta(\mathbf{x})$ depending on the parameters stored in θ . Given an input $x \in \mathcal{X}$, the network yields $p_\theta(Y|x)$ as a predictive distribution. In more detail, $\hat{\pi}_k(x) := p_\theta(Y = k|x)$ returns the estimated probability of x belonging to class k . We furthermore denote the confidence of the network via $\hat{p} = \max_k \pi_k(x)$ and the corresponding class prediction as $\hat{y}$. The core principle of CP for classification lies within the definition of a prediction set $\mathcal{C}(X)$. This prediction set can contain one or more labels from $\mathcal{Y}$, as opposed to the single-class prediction $\hat{y}$, or can be empty. The CP framework now theoretically guarantees a coverage guarantee of its prediction sets on the unseen test set, meaning that the set of true test labels Y_{test} lies within the (set of) prediction sets $\mathcal{C}(X_{\text{test}})$ within a certain probability

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1 - \alpha, \quad 0 < \alpha \ll 1. \quad (3)$$

Here, α denotes a prespecified confidence level, which steers the error rate of the coverage guarantee. To construct the prediction sets, we set aside a calibration set different from the training and testing sets. The weak assumption mentioned earlier for the coverage guarantee to hold is the *exchangeability* [10] of those three sets. Concretely, any finite permutation of a subset of the union of the three sets has the same joint probability distribution as the initial subset. As a next step, we calculate *conformal scores* $s(x, y) \in \mathbb{R}$ on the calibration set of size n_{calib} that supposedly quantify the likeliness of the model's prediction. Based on the scores calculated from the calibration set, a threshold is set via $\hat{t} := ((n_{\text{calib}} + 1)(1 - \alpha))/n_{\text{calib}}$. Then, we utilize the exchangeability assumption to carry over this likelihood behavior to the unseen test data. Concretely, we derive the predictive scores on the test set and define the prediction sets as follows:

$$\mathcal{C}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{t}\}. \quad (4)$$

The following versions of CP for classification problems with neural networks solely differ in their choice of conformal scores. For the *LAC* [29] algorithm, we set $s(x, y_{\text{true}}) = 1 - \hat{\pi}_{y_{\text{true}}}(x)$, namely one minus the softmax value of the true class, which we know for the calibration set. The threshold $\hat{t}$ is then calculated as defined above and denotes the remaining uncertainty for true class predictions in $(1 - \alpha)\%$ of all calibration data points. This uncertainty is taken as the upper bound for (the maximum of) the predicted softmax probabilities on the test set. Due to the assumed exchangeability, the coverage guarantee of $(1 - \alpha)\%$ is fulfilled.

As an adoption of this method, [30] introduced *APSs*. The goal here lies in optimizing the conditional coverage, referring

to coverages within certain classes or subgroups. The authors propose as a score function

$$s(x, y_{\text{true}}) = \sum_{j=1}^{k=y_{\text{true}}} \hat{\pi}_{\text{ord},j}(x) \quad (5)$$

which includes all ordered softmax values $\hat{\pi}_{\text{ord},1}(x), \dots, \hat{\pi}_{\text{ord},K}(x)$ from highest to lowest until reaching the ground-truth class (known for the calibration set). Again, the quantile is formed as above, and the prediction set is formed via (4). Building upon *APS*, [31] proposed an improved version termed *RAPSs*. The authors imposed an additional regularization term to enforce even smaller prediction set sizes. With $\lambda \geq 0$ and $k_{\text{reg}} \geq 0$ being the regularization hyperparameters, they changed the scoring function to

$$s(x, y_{\text{true}}) = \sum_{j=1}^{k=y_{\text{true}}} \hat{\pi}_{\text{ord},j}(x) + \lambda (\sigma_x(y_{\text{true}}) - k_{\text{reg}})^+ \quad (6)$$

where $\sigma_x(y)$ denotes the ranking of the true label y_{true} within the softmax pseudo-probabilities. The prediction sets are then formed similarly as before via (4).

Lastly, we consider the *top-k* conformal calibration method, which takes the classes with the k highest softmax values as the prediction set. Note that k gets optimized on the calibration dataset to satisfy the theoretical coverage guarantees. In more detail, k gets chosen as the $\lceil (n_{\text{calib}} + 1)(1 - \alpha) \rceil$ largest value of the ranks of the true classes within the ordered softmax values of the predictions on the calibration dataset.

C. Evaluation Metrics for CP

The overall coverage serves as the key metric for CP. It measures the percentage of prediction sets that actually contain the true label. For an open-world application of CP, the coverage also represents the key promise of the framework: Even if the single-class prediction is incorrect, we can rely on the prediction set to restrict the space of possible classes for a given sample. Typically, coverage and the following metrics are computed during inference on the held-out test set of size n_{test} , for which the measured marginal coverage reads as follows:

$$\text{Cov} := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{y_i \in \mathcal{C}(x_i)\}}. \quad (7)$$

The coverage is denoted as marginal, since the summation is over the entire test dataset and not class-specific. Ideally, the measured marginal coverage fulfills the theoretical marginal coverage guarantee and additionally produces small prediction sets. The average prediction set size, also termed *inefficiency* [72], is used as a metric for this. On the test set, it is calculated via

$$\text{Ineff} := \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(x_i)|. \quad (8)$$

As a peculiarity of the studied dataset, we furthermore investigate the degree to which the different CP algorithms cover the votes cast by the human experts during the labeling stage in their prediction sets. For this, fix $i \in \{1, \dots, n_{\text{test}}\}$ and test input image x_i and define $V_i := \cup_{j=1}^M \{V_{ij}\} \subset \{1, \dots, K\}$ as

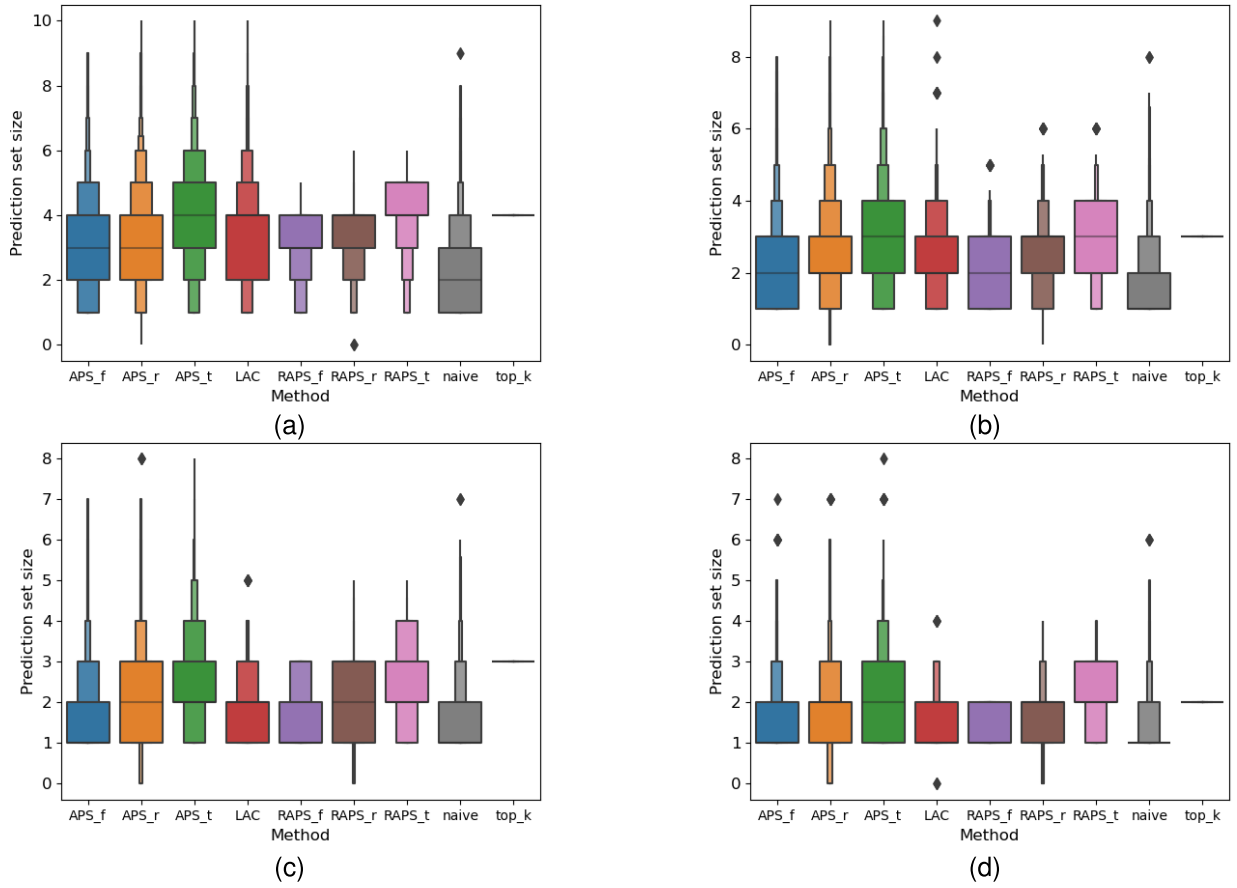


Fig. 2. Prediction set sizes by values of α for the different conformalization methods. x-axis: CP method, y-axis: prediction set size. (a) $\alpha = 0.05$. (b) $\alpha = 0.1$. (c) $\alpha = 0.15$. (d) $\alpha = 0.2$.

the set of classes that experts voted for. We then look at the coverage of all labels that received votes, denoted by

$$\text{Cov}_{\text{lab},i} := \frac{|V_i \cap \mathcal{C}(x_i)|}{|V_i|}. \quad (9)$$

Although the label votes coverage yields the percentage of all cast label votes that are covered by the prediction set, it does not penalize large prediction sets, which naturally include more labels voted. For this reason, we also investigate the *Intersection over Union (IoU)* between the two sets, namely the prediction set and the set of labels voted by the experts

$$\text{IoU}_{\text{pred.-vot},i} := \frac{|V_i \cap \mathcal{C}(x_i)|}{|V_i \cup \mathcal{C}(x_i)|}. \quad (10)$$

Neither metrics yet take into account the actual number of votes cast per class. In order to specifically include this information, we modify the above metrics to focus on the realizations of Y_{ik} instead of just the support. Again, we can first investigate the coverage of the label votes by measuring the percentage of all votes covered by the prediction set

$$\text{Cov}_{\text{lab.-vot},i} := \frac{\sum_{k=1}^K Y_{ik} |V_i \cap \mathcal{C}(x_i) \cap \{k\}|}{\sum_{k=1}^K Y_{ik} |V_i \cap \{k\}|}. \quad (11)$$

For the IoU, the two terms of the fraction can be investigated separately. For the intersection, we can use the above formulation, which measures all votes cast that lie in the prediction set. For the union, we need to consider how to treat classes

that are in the prediction set and have no votes cast by the experts. Since we do not rank the members of a prediction set by likelihood, we assign equal probability to all members and use this uniformity in the calculation of the union. Hence, the IoU considering all individual votes reads as follows:

$$\begin{aligned} \text{IoU}_{\text{lab.-vot},i} &:= \frac{\sum_{k=1}^K Y_{ik} |V_i \cap \mathcal{C}(x_i) \cap \{k\}|}{\sum_{k=1}^K (Y_{ik} |(V_i \cup \mathcal{C}(x_i)) \cap \{k\}| + |(\mathcal{C}(x_i) \setminus V_i) \cap \{k\}|)}. \end{aligned} \quad (12)$$

IV. CP EXPERIMENTS

A. Coverage and Inefficiency Analysis

First, we compare the prediction set sizes produced by the different CP algorithms introduced in Section III-B for the building class test samples. In the following, *naive*, *top_k*, and *LAC* describe the respective methods in the visualizations, while *APS* and *RAPS* can be further distinguished. For these two methods, one can decide whether the label that exceeds the threshold is always included (ending *_t*), is always excluded (ending *_f*), or is decided randomly (ending *_r*). Note that solely the random configuration achieves the exact marginal coverage guarantee theoretically. Fig. 2 displays the range of the prediction set sizes of the test set for the different methods, separately for different chosen confidence levels α . Due to the nature of conformalization, lower α values

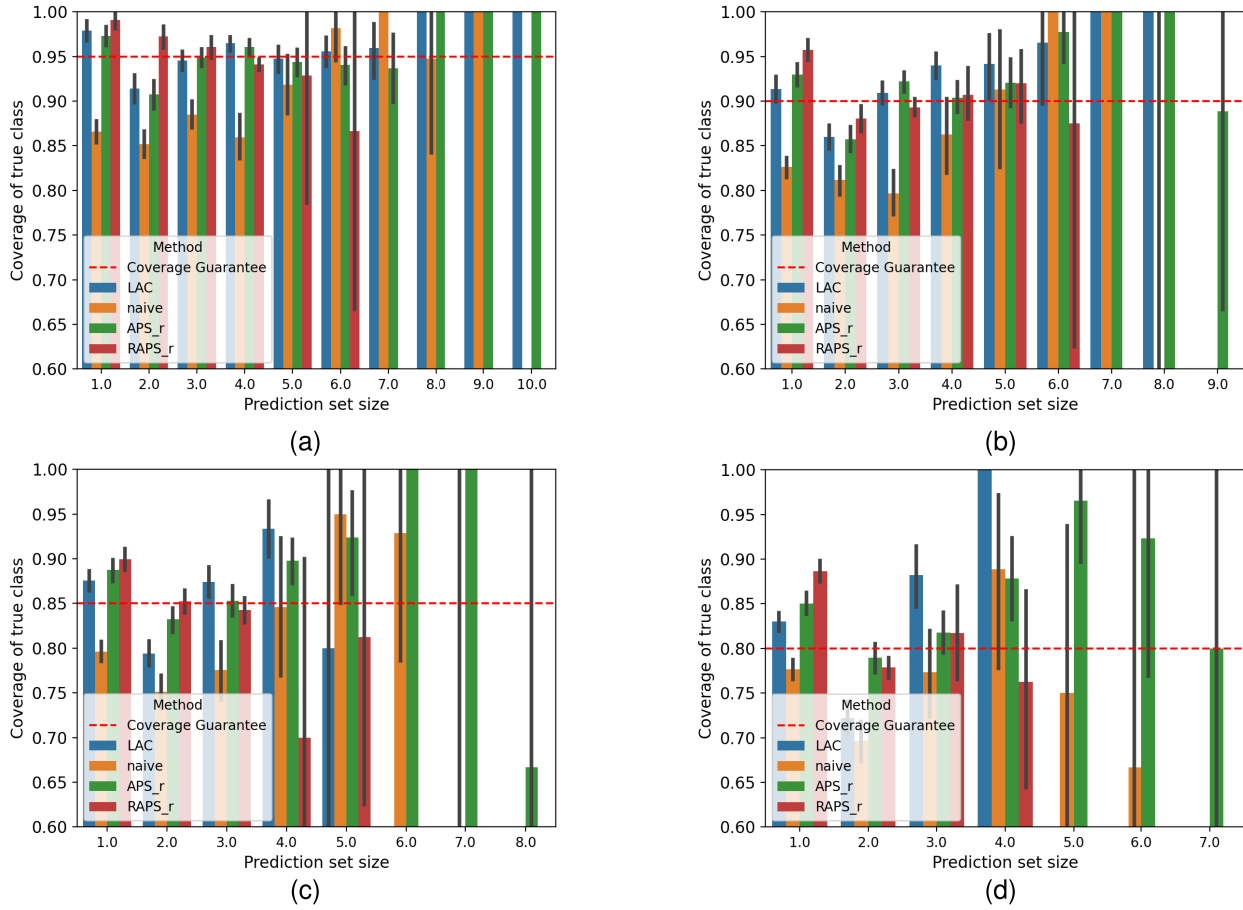


Fig. 3. Marginal coverages of chosen (conformalized) predictors grouped by prediction set sizes, separately for different α values. x-axis: prediction set size, y-axis: coverage of true class in %. (a) $\alpha = 0.05$. (b) $\alpha = 0.1$. (c) $\alpha = 0.15$. (d) $\alpha = 0.2$.

force the conformalized model to produce larger prediction sets to achieve the desired coverage. At first sight, the naive method can be overlooked. This is despite the fact that the method does not incorporate any conformalization step, but rather sums up the ranked softmax probabilities until the threshold $1 - \alpha$ is reached. On average, it yields the smallest prediction set sizes when evaluated on the test set. This hints at a relatively large degree of confidence, as most often, one or two class softmax pseudo-probabilities are sufficient to reach the desired threshold. Furthermore, note that there exists no theoretical guarantee about the coverage of the true label within the prediction set; hence, the softmax pseudo-probabilities cannot be treated as trustworthy estimates of the coverage probabilities.

At the same time, the spread of the set sizes predicted by the naive method is relatively high. APS methods and the LAC conformalizer yield a similar spread (although APS has more outliers), while the RAPS methods are less widely spread in their results. This underscores the importance and effectiveness of regularizing the prediction set sizes to have reasonable set sizes even in cases where high uncertainty exists in the softmax pseudo-probabilities. Note, however, that the random versions of APS and RAPS yield empty prediction sets in some cases, which is undesirable. For $\alpha = 0.15$, which represents a realistic scenario by matching the observed label confidence of the dataset with the desired coverage guarantee, the RAPS_f

method yields superior performance as it matches closely with the naive method in terms of average prediction set size while showing a smaller spread of the sizes.

Inefficiency by itself, however, does not adequately describe the quality of the predictive uncertainty of a (conformalized) predictor. The theoretical coverage guarantee of the conformalization environment should also be marginally met. To investigate this in more detail, we first look at the marginal coverage (9) for each method for varying α values given the prediction set size in Fig. 3. Note that the _t and _f versions of APS and RAPS were discarded for the latter since they do not fulfill the marginal coverage guarantees in theory. Furthermore, the top-k method was intentionally left out due to prediction set sizes far above average. Fig. 3 displays the coverages as a barplot with the standard error as a grey bar for the leftover methods by prediction set size (separately for each α value).

Since the coverage guarantee is only given marginally, it is of no surprise that even for the conformalized methods, the guarantee per prediction set size is not always met. Yet, the error bar crosses the coverage line in almost any case. In this investigation, now the clear downside of the unconformalized naive method becomes apparent: Although often producing very small prediction sets, the marginal coverage is often significantly below the theoretical coverage guarantee. The randomized versions of APS and in particular of RAPS can

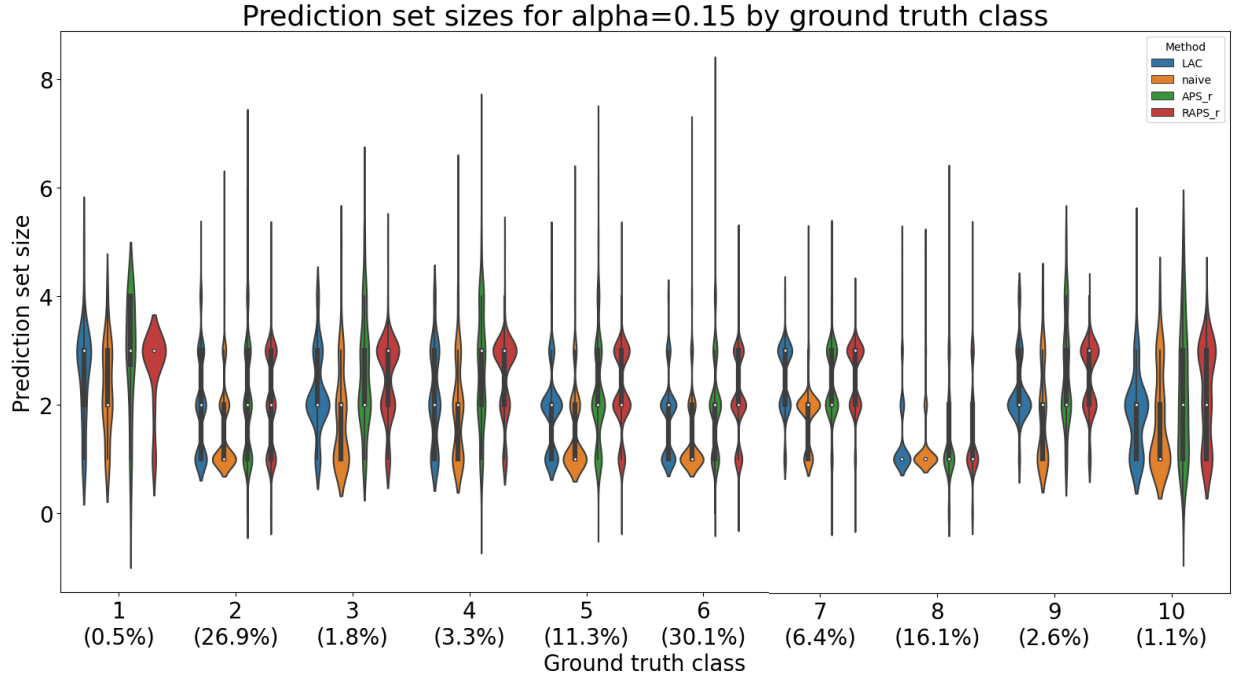


Fig. 4. Inefficiencies per ground-truth class for all methods. Relative frequencies of the classes are listed below the respective x-axis ticks. x-axis: LCZ building class (relative frequency in parentheses in %), y-axis: prediction set size.

hereby glance, underscoring the effectiveness of the produced predictive uncertainties by means of the prediction set sizes.

Even though the coverage guarantee of the proposed conformalizers only holds marginally, it is still of great value to investigate the inefficiencies on a class level. To do so, we can look at the inefficiencies of all methods grouped by the underlying ground truth class (see e.g. Fig. 4 for $\alpha = 0.15$). Unsurprisingly, the different (conformalized) methods yield similar patterns of the prediction set sizes on class level, only differing in their spread. Yet more interestingly, underrepresented building classes do not necessarily yield larger prediction set sizes. Seemingly, some classes are associated with higher uncertainty, regardless of their share in the dataset.

In more detail, we show the observed coverages of the true class for every method given a varying degree of sensitivity in Fig. 5. The randomized APS and RAPS methods here lie almost exactly on top of the ideal diagonal line (see Fig. 5). Note that in particular for the prediction set size of one, the marginal coverage coincides with the accuracy and furthermore gives a notion of calibration for the naive method: Since the confidence (i.e. highest softmax probability) in this setting lies at least at the threshold $1 - \alpha$, a well-calibrated model should match this value with its accuracy. Here, an apparent mismatch is visible for the naive method, hinting at an overconfident baseline model. Note that the margin is smaller for higher α values, hence the overconfidence is particularly pronounced in the very high confidence regions. After conformalization, the calibration drastically improves. Thus, prediction sets of size one become more trustworthy and better calibrated. However, they become less likely because conformalization tries to balance overconfidence and fulfilling the marginal coverage guarantee.

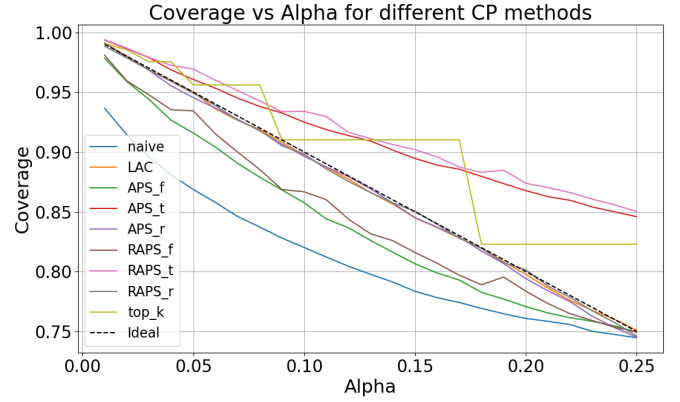


Fig. 5. Marginal coverages of all considered predictors for varying α values. x-axis: α value, y-axis: coverage of true class in %.

B. CP Uncertainty as Indicator

In contrast to many other UQ methods, CP offers a very clear and precise notion of predictive uncertainty via the prediction set size. Naturally, this uncertainty includes both aleatoric and epistemic uncertainty. The hope of conformalization lies in refining a classifier's uncertainty. While coverage checks the validity of the prediction set size, the question can be raised whether it can also successfully hint at other uncertainty quantities. In the following, we therefore want to investigate its indicative power with respect to the underlying label uncertainty measured by the expert entropy and to a notion of epistemic uncertainty by means of predictive performance.

We first start by comparing the prediction set size with the underlying degree of aleatoric label uncertainty, which we measure via the Shannon entropy of the soft label (2) defined

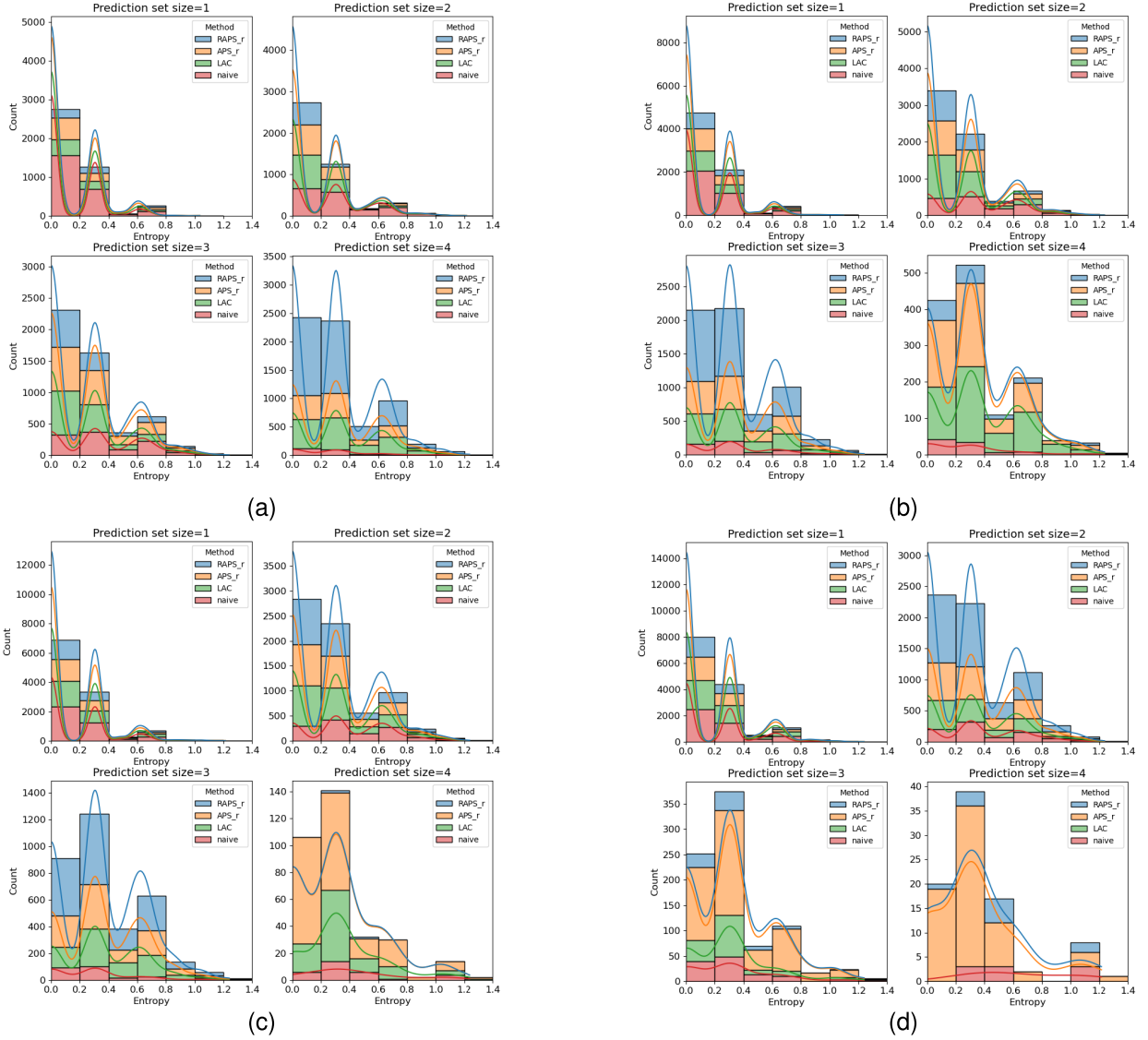


Fig. 6. Label uncertainty (by means of Shannon entropy) grouped by prediction set size, separately for each α value. x-axis: Entropy of expert votes, y-axis: Number of observed instances. (a) $\alpha = 0.05$. (b) $\alpha = 0.1$. (c) $\alpha = 0.15$. (d) $\alpha = 0.2$.

via

$$H(\mathbf{y}_{\text{distr},i}) = - \sum_k \mathbf{y}_{\text{distr},i,k} \cdot \log \mathbf{y}_{\text{distr},i,k}. \quad (13)$$

We argue that while other sources of aleatoric uncertainty exist and cannot be directly measured, the label uncertainty (as part of aleatoric uncertainty) can be estimated by the observed expert entropy. Although our estimate of label uncertainty is only based on ten expert votes, we claim that the diversity of the observed degrees of label uncertainty and the sheer size of the dataset justify the validity of the following experiments. Visually, we inspect the interdependency between the total uncertainty of the prediction sets and the measured aleatoric uncertainty of the labels in Fig. 6 on the test set of the building-related LCZ classes. For each prediction set size separately, we look at the distribution of the entropies in a stacked barplot. Note first of all, as was already observed in

previous experiments, that the unconformalized naive method yields generally lower prediction set sizes. What also becomes apparent is the larger share of medium and high entropy values for the conformalized methods, e.g., in the $\alpha = 0.05$ setting. The effect is less visible for higher α values since the overall average prediction set size decreases. Yet, the trend is still visible as the red bars of the naive method are diminishing for higher prediction sets.

To better answer the question of whether higher prediction sets of conformalized methods can possibly hint at larger underlying label uncertainty (estimated by expert entropy), we state the conditional probabilities of the presence of label uncertainty given a fixed prediction set size in Table I above. For the methods previously discussed, the table presents the observed probabilities of a sample in the test set of the building-related LCZ classes having no degree of label uncertainty as well as the complementary probability of having label

TABLE I
INEFFICIENCY ANALYSIS FOR THE MOST PROMISING CP METHODS (RAPS AND APS IN RANDOMIZED VERSION), FOR $\alpha = 0.05$

α	Method	Ineff.	$P(H(y) = 0 \text{Ineff} = k)$	$P(H(y) > 0 \text{Ineff} = k)$	Odds($H(y) > 0$)
0.05	APS	1	66.39 %	33.61 %	0.51
		2	62.91 %	37.09 %	0.59
		3	42.74 %	57.26 %	1.34
		4	38.37 %	61.63 %	1.61
		5	39.16 %	60.84 %	1.55
		6	53.24 %	46.76 %	0.88
		7	40.51 %	59.49 %	1.47
		8	12.50 %	87.50 %	7.00
		9	22.58 %	77.42 %	3.43
0.05	LAC	1	61.33 %	38.67 %	0.63
		2	64.57 %	35.43 %	0.55
		3	50.58 %	49.42 %	0.98
		4	35.25 %	64.75 %	1.84
		5	39.65 %	60.35 %	1.52
		6	45.79 %	54.21 %	1.18
		7	49.66 %	50.34 %	1.01
		8	24.32 %	75.68 %	3.11
		9	13.33 %	86.67 %	6.50
		10	16.67 %	83.33 %	5.00
0.05	RAPS	1	51.17 %	48.83 %	0.95
		2	82.66 %	17.34 %	0.21
		3	54.00 %	46.00 %	0.85
		4	41.57 %	58.43 %	1.41
		5	14.29 %	85.71 %	6.00
		6	6.67 %	93.33 %	14.00
0.05	naive	1	60.94 %	39.06 %	0.64
		2	42.45 %	57.55 %	1.36
		3	34.89 %	65.11 %	1.87
		4	46.35 %	53.65 %	1.16
		5	48.16 %	51.84 %	1.08
		6	16.36 %	83.64 %	5.11
		7	27.27 %	72.73 %	2.67
		8	10.53 %	89.47 %	8.50

uncertainty, both conditional on varying prediction set sizes (all based on the relative frequencies observed on the hold-out test set). The last column furthermore states the resulting odds of observing label uncertainty for a given α , method, and prediction set size, given by

$$\text{Odds}(H(y) > 0) := \frac{P(H(y) > 0 | \text{Ineff} = k)}{P(H(y) = 0 | \text{Ineff} = k)}. \quad (14)$$

First of all, we observe that generally, across methods, a higher prediction set size increases the likelihood (expressed by the odds) of the presence of label uncertainty (measured via expert entropy). This likelihood differs from method to method: While for the naive method, the odds are relatively small for prediction sets smaller than 5, the odds for the APS method are already quite large beginning with prediction set size 3 and higher (with the exception of prediction set size 6, which might have a small support). Interestingly, the odds for the RAPS method are only starting to drastically increase after prediction set size 3. Therefore, the indication for this method lags a bit behind the other methods. The LAC method, on the other hand, can be better suited for this task and presents large odds for prediction sets of size 4 and larger (again with

a minor exception of prediction set size 7 due to insufficient data support).

To understand on a more fine-grained level to what extent the labels voted by the ten experts are covered by the prediction sets produced by the various conformalizers, we investigate the two quantities introduced in (11) and (12). The results are combined and displayed in Fig. 7. First, note the small differences across methods for the smallest investigated α value. Yet, there still exists a significant gap in label votes coverage between the unconformalized *naive* method and the conformalized methods. This gap is observed across the different values of α , indicating a lack of indication for label uncertainty for the unconformalized method. Unsurprisingly, the overall numbers decrease for large α values, since this leads to smaller prediction set sizes in general. As expected, the conformalized configurations of APS and RAPS, which always include the borderline label, achieve higher coverages than the configurations that never include this label or which decide randomly (the latter being the only case where the coverage guarantee is exactly met). More interestingly, the IoU is also positively impacted by the inclusion of the borderline label, in particular for larger α values. Hence, one can conclude that the borderline label often carries valuable

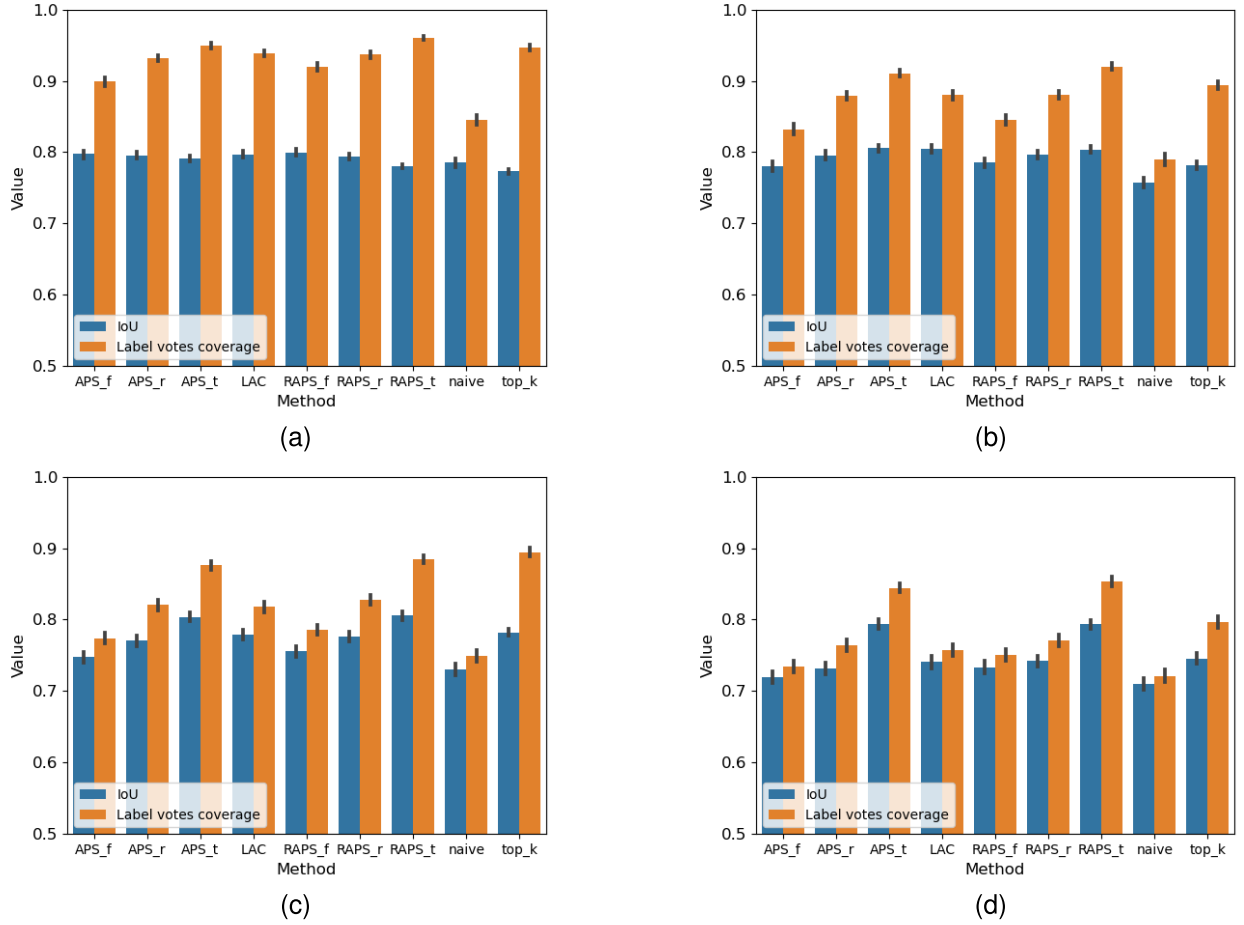


Fig. 7. Label votes coverages and IoUs by values of α . x-axis: CP method, y-axis: IoU or coverage value in %. (a) $\alpha = 0.05$. (b) $\alpha = 0.1$. (c) $\alpha = 0.15$. (d) $\alpha = 0.2$.

information about the intrinsic label uncertainty (measured among the ten experts). It is noteworthy here that the overall highest IoU is already achieved at the $\alpha = 0.15$ level, hence forcing larger prediction sets with even smaller α values is counterproductive for indicating label uncertainty (by means of expert entropy).

As a further analysis, the interdependency between prediction set size and epistemic uncertainty was studied. We investigate this dependency by looking at the accuracy of the classifier with respect to the prediction set size. Underlying this is the claim that an uncertain model is more often incorrect in its prediction than a certain one (in particular if it is reasonably well calibrated). Although the theoretical coverage guarantee makes no statement about the accuracy, we would still expect the conformalized model to have higher accuracy at predictions with a smaller prediction set size. Consequently, a large prediction set size should, if expressing higher uncertainty correctly, yield a lower accuracy. This uncertainty expressed by the conformalized model is then compared to the epistemic uncertainty reflected by the accuracy. The results are presented in Fig. 8. One can note that especially for small prediction set sizes, the conformalized models always yield higher accuracies; hence, predictions from these models have become more trustworthy through

conformalization. In conclusion, even though the naive method produces the smallest prediction sets on average, the accuracy of a given prediction set size is almost always higher after conformalization. Therefore, the conformalizer's uncertainty yields a better indicator for the model's uncertainty than the uncertainty from an unconformalized model.

V. AL EXPERIMENTS

The overall goal of the proposed AL experiments is to check the quality of different UQ methods. If the uncertainties are valid, adding uncertain (i.e., expressive) samples helps the model generalize to unseen test data better. Hence, a fast convergence of the test performance metrics is desired. In particular, the generalization performance should exceed the random baseline, which increases the training pool with randomly queried data. The uncertainties are used to iteratively query new data (with high predictive uncertainties), which are then added to the initial training dataset and labeled by the human oracle. The model is then retrained on the new pool of training data before uncertainties are used to query data again. We verify the quality of the uncertainties by measuring the performance of the model on the hold-out test set after every loop.

We first briefly describe the general AL loop. An initial training set is formed via a random subset of the entire

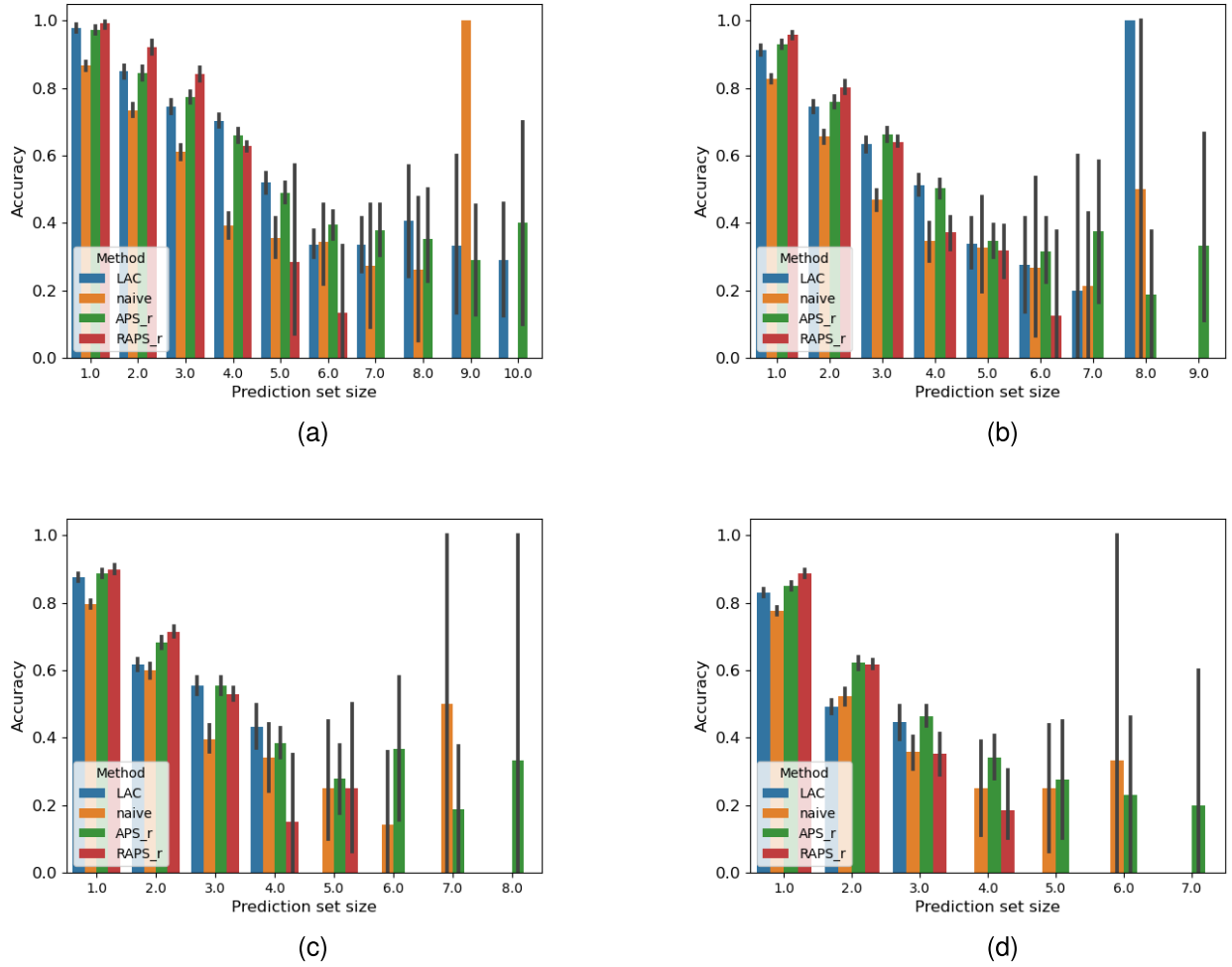


Fig. 8. Accuracy grouped by prediction set size for (conformalized) predictors, separately for different α values. x-axis: prediction set size, y-axis: accuracy in %. (a) $\alpha = 0.05$. (b) $\alpha = 0.1$. (c) $\alpha = 0.15$. (d) $\alpha = 0.2$.

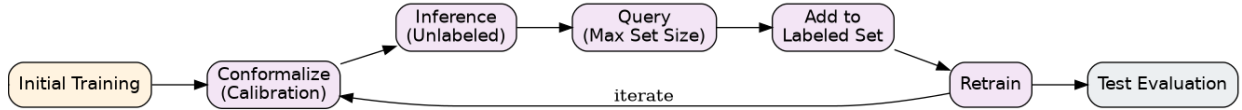


Fig. 9. Schematic CP-AL workflow.

available training data. Next to the usual test set, a calibration set is formed, where exchangeability should be guaranteed between the three separate sets. After initializing a model and training it on the initial training set, we use the model to predict the remaining training data. Based on these predictions, and using multiple querying strategies, we select data points from the remaining training pool that get added to the initial training data. After that, the model is retrained on the increased training data pool and again used for inference. Also, at this stage, we evaluate the model's test performance and store it for later.

As a novelty, we introduce inefficiency-based AL. While previous works such as [46] incorporated the score function of a conformalizer for querying, we now shift the focus toward the estimated prediction sets. We show a schematic workflow

in Fig. 9. In particular, this inefficiency-based approach does not directly depend on the softmax pseudo-probabilities anymore, which might reflect over- or underconfidence of the underlying model. Therefore, before each inference step, we conformalize our classifier and utilize the resulting threshold for estimating the prediction sets for the remaining training data. Then, we query the points that get added to the training set with respect to the prediction set size. Since large prediction sets indicate a high degree of uncertainty in the prediction, we add those points to our training set to facilitate the model's generalization ability on the test set. We test this with a slightly shifted interest: While still evaluating the test accuracy and ensuring the theoretical coverage guarantee of the true label, the focus now additionally lies more on lower and more expressive predictive uncertainty.

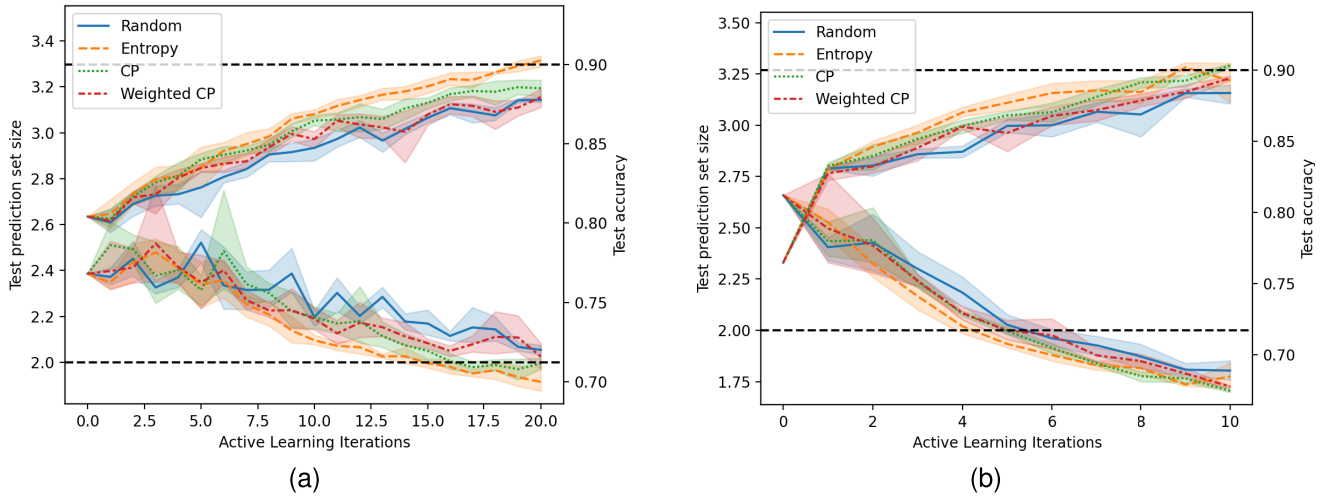


Fig. 10. Random split setting with varying no. of samples per AL iteration. x-axis: AL iteration, y-axis: test prediction set size (left) and test accuracy in % (right). (a) 500 samples per AL iteration. (b) 1000 samples per AL iteration.

This is measured via the average prediction set size on the test set. Suppose inefficiency-based AL leads to a faster convergence of the average test prediction set size (to, e.g., a lower bound). In that case, we can deduce that uncertainty-aware AL using CP can generate lower predictive uncertainties while guaranteeing a solid level of coverage of the true label. Suppose, furthermore, or instead, the inefficiency-based AL leads to a faster increase in test accuracy. In that case, we can deduce that adding more uncertain samples to the training pool helps the model to adapt more quickly to unseen data.

As described earlier, we now test the quality of the predictive uncertainty by means of an AL downstream task. For this, we randomly split the entire pool of available data into a large training pool (70%) as well as reasonably sized calibration and test data pools (15% each). Out of the large training pool of approximately 35k satellite patches, we randomly select 5k samples, which serve as an initial training pool. In contrast, the remaining samples are denoted as the unlabeled training pool. A randomly initialized model is then trained on the 5k training samples (pretraining stage). In each AL iteration, the separately trained models predict on the unlabeled training pool and select a fixed number of data points through their respective querying strategies. As a novelty of this work, we investigate the predictive uncertainty of conformalization methods as a querying strategy. We choose the *RAPS* conformalizer with $\alpha = 0.05$, with the inclusion of the borderline label to ensure sufficiently large prediction set sizes. After each training step (and initially after the pretraining stage), the model is then conformalized by means of the hold-out calibration dataset, and the prediction sets are computed for the unlabeled training pool. These prediction sets are then used for querying in two distinct ways: In the first setting, the points are selected solely based on their prediction set sizes (*CP*). Meanwhile, in a second setting, the prediction set size gets weighted with the inverse class frequencies of the set members (*weighted CP*) to foster the inclusion of underrepresented classes. We note at this point that the conformalization incurs an additional computational cost, as shown in Table II. Concluding, an overhead of ca. 25% can be

TABLE II
MEAN TIMINGS OVER ALL AL LOOPS FOR 1 EPOCH OF TRAINING

Metric	Avg. Time (sec)
Conformalization	5.68
Prediction + Querying	10.67
Training	12.83

attributed to conformalization. Note, however, that in practice we have run multiple epochs per AL Loop, hence the overhead is diminishingly small.

Next to a baseline model, which samples the points that get added to the training data randomly, we also investigate softmax entropy-based querying. This technique is widely used in the machine learning community [73], [74], [75] and utilizes the Shannon entropy of the predicted softmax vector for querying (higher is preferred, which supposedly represents higher predictive uncertainty). The promise of AL lies in the fast generalization ability of a model once presented with predominantly uncertain samples. Therefore, we limit ourselves in the following experiments to adding a total of 10k samples to the initial training dataset, corresponding to roughly one-third of the available unlabeled training pool. In two separate scenarios, 500 and 1000 samples are added to the training data per step, respectively. We present the results in Fig. 10. In particular, test accuracy and, more importantly, prediction set sizes for the test set predictions are reported in each iteration for five models per configuration (mean $\pm$ standard deviation), each randomly initiated with a different seed.

Multiple conclusions can be drawn from these results. First and foremost, we see clear advantages of the uncertainty-based querying methods over random querying, which justifies the methodological framework and computational overhead (in the case of conformalization). At the same time, we see equal or, in certain cases, even slightly better test performance of uncertainty-based models for a moderate share of data used (ca. 40%) compared to the full model, which gets trained on all

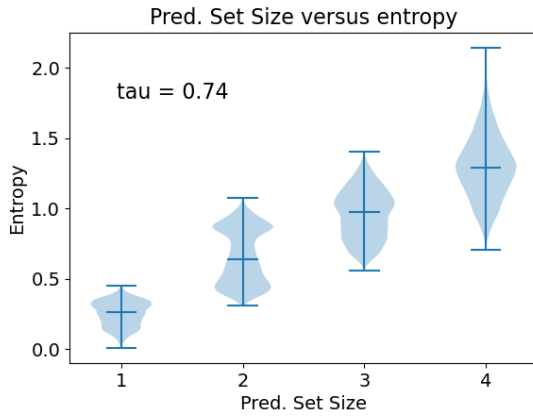


Fig. 11. Softmax entropies and prediction set sizes for the unlabeled training pool. x -axis: prediction set size, y -axis: entropy of expert votes.

available training data right away. Its performance is indicated by the black dashed line. Furthermore, the overall convergence for the higher number of samples per iteration appears to be smoother. Yet for this setting, a slightly higher variation exists overall across the different models for each configuration, in particular for the *CP* and weighted *CP* configurations.

Regarding the superior performance of the entropy-based querying, we investigated the following: Since the prediction set size is limited to whole positive numbers and bounded from above by the number of classes, it is more tightly bound than the entropy of the softmax. The latter is also bounded from above and below, but can take any value in between the bounds on a continuous scale. We visualized the realizations of the two quantities for the first prediction of the pretrained model on the remaining unlabeled training data pool in Fig. 11. Note the core message of this plot, namely that querying points with respect to maximum entropy will include points solely with the maximum prediction set size, but not vice versa. This is due to the large spread of the entropies given the highest prediction set size.

Lastly, we try to explain the slightly underperforming weighted *CP* method. Note first that the prediction set gives no internal ordering to the classes it contains. Therefore, solely when a prediction set contains an underrepresented class, the weight will drastically increase, and the sample is more likely to be queried, regardless of whether the underrepresented class is the true class or not. Note, however, that the ground-truth label is not known for the unlabeled training pool until the samples are added to the initial training data. Therefore, we cannot utilize this information at the querying stage. Second, the class imbalance is quite large in the dataset, leading to huge weights for underrepresented classes. This can, of course, be targeted and optimized, but lies outside the scope of this work.

One advantage of the weighted *CP* approach becomes apparent in Fig. 12, where now the balanced accuracy (averaging all ten class-specific accuracies with equal weight) is investigated. The balanced accuracy is particularly useful for nonuniform class cardinality yet uniform class importance [76], as is the case for the present dataset. We observe similar trends in predictive uncertainty, but a clear advantage of the weighted *CP* approach over the regular *CP* and the random approach

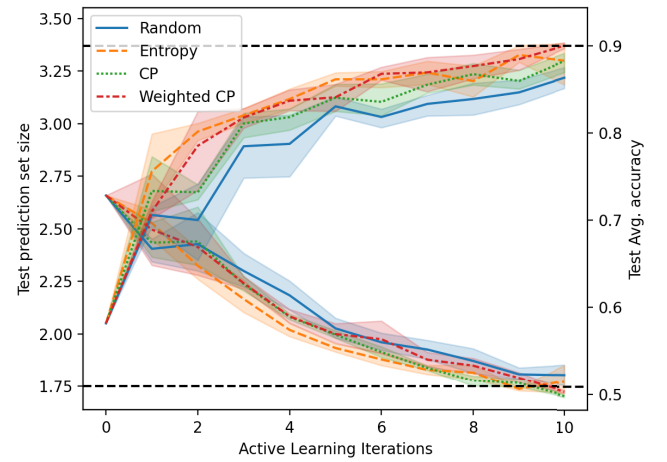


Fig. 12. Random split setting with 1000 samples per AL iteration. x -axis: AL iteration, y -axis: test prediction set size (left) and balanced test accuracy in % (right).

is visible for the balanced accuracy. Therefore, the weighting can be seen as a valuable add-on to the *CP*-based querying, which specifically targets balanced accuracy across all classes.

Note that the gap to the entropy-based querying is again marginal, underscoring the advantage of the entropy-based method discussed earlier. We want to emphasize that the performance gain in balanced accuracy does not sacrifice any loss in overall accuracy, since the models in Figs. 10(a) and 12 are identical. Merely the shift in focus on the balanced accuracy shows the competitive advantage of a weighted *CP*-querying approach for imbalanced remote sensing datasets. These imbalances occur often in practice; hence, a weighted approach of some sort is desirable.

VI. CONCLUSION

We have studied the framework of *CP*, which constructs prediction sets consisting of multiple classes instead of a single class for a given classifier by means of statistical properties. These prediction sets then fulfill theoretical guarantees about the inclusion of the ground-truth label and provide more insights into the model's predictive uncertainty. We have utilized this UQ toolbox for the large-scale land cover classification dataset *LCZ42* (limited to the building-related classes), for which we have multilabel data coming from ten independent expert annotators. As a baseline for comparison, we define an unconformalized method of constructing prediction sets simply by adding up softmax probabilities (highest to lowest) until the desired inclusion is achieved. This method is then compared to more sophisticated conformalized methods such as *LAC* or *RAPS*. Although the baseline often yields smaller prediction sets overall, the coverage guarantee cannot always be fulfilled, and the accuracy with regard to the prediction set size is not as trustworthy as for the conformalized methods. Adding to that, the conformalized methods also manage to cover more of the intrinsic label uncertainty (estimated by the expert entropy), hence presenting a more complex notion of predictive uncertainty. Both findings support the hypothesis that the conformalizer's predictive uncertainty presents a good indicator for both aleatoric uncertainty, measured by the intrinsic label uncertainty via expert

entropy, and epistemic uncertainty, under the assumption that higher model uncertainty should, on average, align with lower predictive uncertainty.

As a further study, we evaluated the CP framework by means of AL for all 17 LCZ classes. By utilizing the prediction set size as a query function, we managed to beat the random baseline and to achieve test performance equivalent to that of a full model, using only a fraction of the data, while producing significantly smaller prediction sets compared to the full model. Note that entropy-based querying could even beat the CP-based querying, which underscores the importance of uncertainty in the AL framework. Overall, we conclude that CP offers a promising toolkit for expressing a classifier's predictive uncertainty in an explainable manner. We could show that a conformalized classifier can reach better coverage as well as more trustworthy statements about the uncertainty associated with its prediction.

We also see huge potential for CP in the broader field of modeling remote sensing data, which often requires an explainable and trustworthy notion of predictive uncertainty. While the transferability of the proposed method to other geographical regions highly depends on the exchangeability, the uncertainty in the form of a prediction set greatly improves its explainable power and expressiveness. Regarding the applicability of CP to different remote sensing tasks, we are generally limited to classification and regression tasks due to its construction. Whenever users benefit from prediction sets or intervals as opposed to class or point predictions, the advantage of CP-based approaches becomes obvious. Hence, applications to land cover classification or segmentation tasks, biomass estimation, harvest readiness classification, and many more are feasible. We see possible future research in the direction of utilizing multiseasonal or multitemporal remotely sensed data, as this greatly improves the data foundation for the conformalizer. Furthermore, self-supervised methods for representation learning would greatly benefit from CP integration as it would act as an internal calibration module.

REFERENCES

- [1] D. Landgrebe and E. Malaret, "Noise in remote-sensing systems: The effect on classification error," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-24, no. 2, pp. 294–300, Mar. 1986.
- [2] K. Hechinger, X. X. Zhu, and G. Kauermann, "Categorising the world into local climate zones: Towards quantifying labelling uncertainty for machine learning models," *J. Roy. Stat. Soc. Ser. C, Appl. Statist.*, vol. 73, no. 1, pp. 143–161, Jan. 2024.
- [3] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.
- [4] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6405–6416.
- [5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017. [Online]. Available: <https://openreview.net/forum?id=Hkg4TI9xl>
- [6] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7498–7512.
- [7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, vol. 29. Cham, Switzerland: Springer, 2005.
- [8] V. Vovk, "Transductive conformal predictors," in *Proc. 9th IFIP WG Int. Conf. Artif. Intell. Appl. Innov. (AIAI)*, Paphos, Cyprus. Cham, Switzerland: Springer, Sep. 2013, pp. 348–360.
- [9] H. Papadopoulos, V. Vovk, and A. Gammerman, "Conformal prediction with neural networks," in *Proc. 19th IEEE Int. Conf. Tools With Artif. Intell. (ICTAI)*, Oct. 2007, pp. 388–395.
- [10] R. Evans, "Predicting the unpredictable," *Astron. Geophys.*, vol. 38, no. 6, p. 4, Dec. 1997.
- [11] L. M. Arjona, P. Bhattacharjee, and P. Jung, "Onboard conformal prediction for domain shift in Earth observation," *Proc. SPIE*, vol. 13670, pp. 95–102, Oct. 2025.
- [12] T. Fritz, T. Gonda, and P. Perrone, "De Finetti's theorem in categorical probability," 2021, *arXiv:2105.02639*.
- [13] L. Zhou, Z. Shao, S. Wang, and X. Huang, "Deep learning-based local climate zone classification using Sentinel-1 SAR and Sentinel-2 multispectral imagery," *Geo-Spatial Inf. Sci.*, vol. 25, no. 3, pp. 383–398, Jul. 2022.
- [14] D. Angluin, "Queries and concept learning," *Mach. Learn.*, vol. 2, no. 4, pp. 319–342, Apr. 1988.
- [15] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorolog. Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.
- [16] V. Kirschner, K. Macku, D. Moravec, and J. Mañas, "Measuring the relationships between various urban green spaces and local climate zones," *Sci. Rep.*, vol. 13, no. 1, p. 9799, Jun. 2023.
- [17] S. Doda et al., "Interpretable deep learning for consistent large-scale urban population estimation using Earth observation data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, Apr. 2024, Art. no. 103731.
- [18] S. J. Quan and P. Bansal, "A systematic review of GIS-based local climate zone mapping studies," *Building Environ.*, vol. 196, Jun. 2021, Art. no. 107791.
- [19] Q. Chunping, M. Schmitt, M. Lichao, and Z. Xiaoxiang, "Urban local climate zone classification with a residual convolutional neural network and multi-seasonal Sentinel-2 images," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens. (PRRS)*, Aug. 2018, pp. 1–5.
- [20] Y. Fu et al., "Mapping impervious surfaces in town–rural transition belts using China's GF-2 imagery and object-based deep CNNs," *Remote Sens.*, vol. 11, no. 3, p. 280, Jan. 2019.
- [21] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, 2023.
- [22] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13153–13164.
- [23] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. ICML*, 2015, pp. 1613–1623.
- [24] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.
- [25] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," 2021, *arXiv:2102.11409*.
- [26] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7047–7058.
- [27] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.
- [28] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021, *arXiv:2107.07511*.
- [29] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *J. Amer. Stat. Assoc.*, vol. 114, no. 525, pp. 223–234, Jan. 2019.
- [30] Y. Romano, M. Sesia, and E. Candès, "Classification with valid and adaptive coverage," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3581–3591.
- [31] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=eNdiU_DbM9
- [32] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Statist.*, vol. 51, no. 2, pp. 816–845, Apr. 2023.
- [33] Z. Liang, Y. Zhou, and M. Sesia, "Conformal inference is (almost) free for neural networks trained with early stopping," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 20810–20851.

- [34] D. Bethell, S. Gerasimou, and R. Calinescu, "Robust uncertainty quantification using conformalised Monte Carlo prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 20939–20948.
- [35] L. Dabah and T. Tirer, "On temperature scaling and conformal prediction of deep classifiers," 2024, *arXiv:2402.05806*.
- [36] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. ICML*, vol. 2, no. 4, 2000, p. 6.
- [37] P. Ren et al., "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021.
- [38] P. Mitra, B. Uma Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognit. Lett.*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.
- [39] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.
- [40] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.
- [41] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.
- [42] J. Lin, T. Yu, and Z. J. Wang, "Rethinking crowdsourcing annotation: Partial annotation with salient labels for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5804112.
- [43] S.-S. Ho and H. Wechsler, "Query by transduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1557–1571, Sep. 2008.
- [44] V. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Generalized query by transduction for online active learning," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops ICCV Workshops*, Sep. 2009, pp. 1378–1385.
- [45] T. Löfström, U. Johansson, and H. Boström, "Effective utilization of data in inductive conformal prediction using ensembles of neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.
- [46] S. Matiz and K. E. Barner, "Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification," *Pattern Recognit.*, vol. 90, pp. 172–182, Jun. 2019.
- [47] S. Matiz and K. E. Barner, "Conformal prediction based active learning by linear regression optimization," *Neurocomputing*, vol. 388, pp. 157–169, May 2020.
- [48] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616819.
- [49] C. Koller, P. Jung, and X. X. Zhu, "Can land cover classification models benefit from distance-aware architectures?" *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [50] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Intl. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1321–1330.
- [51] C. Koller, G. Kauermann, and X. X. Zhu, "Going beyond one-hot encoding in classification: Can human uncertainty improve model performance in Earth observation?" *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5400211.
- [52] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Ruskovskiy, "Human uncertainty makes classification more robust," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9616–9625.
- [53] P. Bhattacharjee and P. Jung, "Onboard out-of-calibration detection of deep learning models using conformal prediction," 2024, *arXiv:2405.02634*.
- [54] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson, "Uncertainty quantification for probabilistic machine learning in Earth observation using conformal prediction," *Sci. Rep.*, vol. 14, no. 1, p. 16166, Jul. 2024.
- [55] D. Valle, R. Izbicki, and R. V. Leite, "Quantifying uncertainty in land-use land-cover classification using conformal statistics," *Remote Sens. Environ.*, vol. 295, Sep. 2023, Art. no. 113682.
- [56] D. Valle, R. Leite, R. Izbicki, C. Silva, and L. Haneda, "Local uncertainty maps for land-use/land-cover classification without remote sensing and modeling work using a class-conditional conformal approach," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 135, Dec. 2024, Art. no. 104288.
- [57] K. Liu, T. Sun, H. Zeng, Y. Zhang, C.-M. Pun, and C.-M. Vong, "Spatial-aware conformal prediction for trustworthy hyperspectral image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 9, pp. 8754–8766, Sep. 2025.
- [58] N. Kakhani, S. Alamdar, N. M. Kebonye, M. Amani, and T. Scholten, "Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction," *Remote Sens.*, vol. 16, no. 3, p. 438, Jan. 2024.
- [59] N. Zhu et al., "Inductive conformal prediction enhanced LSTM-SNN network: Applications to birds and UAVs recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [60] M. Farag, J. Kierdorf, and R. Roscher, "Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 651–659.
- [61] F. Meng and T. Song, "Uncertainty forecasting system for tropical cyclone tracks based on conformal prediction," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123743. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417424006092>
- [62] J. Han, N. Mo, J. Cai, L. Ouyang, and Z. Liu, "Advancing the local climate zones framework: A critical review of methodological progress, persisting challenges, and future research prospects," *Humanities Social Sci. Commun.*, vol. 11, no. 1, pp. 1–18, Apr. 2024.
- [63] K. Hammerberg, O. Brousse, A. Martilli, and A. Mahdavi, "Implications of employing detailed urban canopy parameters for mesoscale climate modelling: A comparison between WUDAPT and GIS databases over Vienna, Austria," *Int. J. Climatol.*, vol. 38, no. S1, pp. e1241–e1257, Apr. 2018.
- [64] J. Hidalgo et al., "Comparison between local climate zones maps derived from administrative datasets and satellite observations," *Urban Climate*, vol. 27, pp. 64–89, Mar. 2019.
- [65] Z. Yan et al., "Comparing object-based and pixel-based methods for local climate zones mapping with multi-source data," *Remote Sens.*, vol. 14, no. 15, p. 3744, Aug. 2022.
- [66] S. Cui, X. Wang, X. Yang, L. Hu, Z. Jiang, and Z. Feng, "Mapping local climate zones in the urban environment: The optimal combination of data source and classifier," *Sensors*, vol. 22, no. 17, p. 6407, Aug. 2022.
- [67] A. Vavassori et al., "A combined remote sensing and GIS-based method for local climate zone mapping using PRISMA and sentinel-2 imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 131, Jul. 2024, Art. no. 103944.
- [68] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.
- [69] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.
- [70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [72] V. Vovk, V. Fedorova, I. Nouretdinov, and A. Gammerman, "Criteria of efficiency for conformal prediction," in *Proc. 5th Int. Symp. Conformal Probabilistic Predict. With Appl. (COPA)*, Madrid, Spain. Cham, Switzerland: Springer, Apr. 2016, pp. 23–39.
- [73] P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 62, no. 1, pp. 145–157, Jan. 2000.
- [74] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.
- [75] Z. Qiu, D. J. Miller, and G. Kesidis, "A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 917–933, Apr. 2017.
- [76] A. Gupta, N. Tatbul, R. Marcus, S. Zhou, I. Lee, and J. Gottschlich. (2021). *Class-Weighted Evaluation Metrics for Imbalanced Data Classification*. [Online]. Available: <https://openreview.net/forum?id=PBfaUXYZzU>