Uncertainty-Aware Flood Segmentation from Sentinel 2 Observations with Conformal Prediction

Ioannis Konidakis[†,*], Klea Panayidou[‡,†], Grigorios Tsagkatakis[†,*], Panagiotis Tsakalides[†,*]
[†]Institute of Computer Science, Foundation for Research and Technology - Hellas, Greece.
[‡] European University Cyprus, Nicosia 2404, Cyprus
[*]Computer Science Department, University of Crete, Greece.

***Abstract*—Flood segmentation using supervised deep learning models like U-Net plays a pivotal role in disaster response by facilitating rapid and accurate identification of flood-affected areas. However, ensuring the reliability of these models' predictions is essential, especially in high-stakes applications. Conformal Prediction (CP) provides statistically valid uncertainty estimates and is increasingly recognized as a robust tool for uncertainty quantification. This paper investigates the performance of the two major approaches in CP, namely Inductive CP and k-fold Cross-Validation CP (CV+), in the context of flood segmentation. By evaluating these methods on a baseline bitemporal U-Net model, we demonstrate that CP can offer critical insights into model confidence. Our findings highlight the limitations of Inductive CP in data-scarce scenarios and underscore the advantages of CV+ in achieving a superior balance between calibration and training data usage. This study highlights the importance of CP techniques in enhancing trustworthiness in flood segmentation models.**

***Index Terms*—Flood segmentation, Conformal Prediction (CP), Inductive CP, Cross-Validation CP, Bitemporal Image Data.**

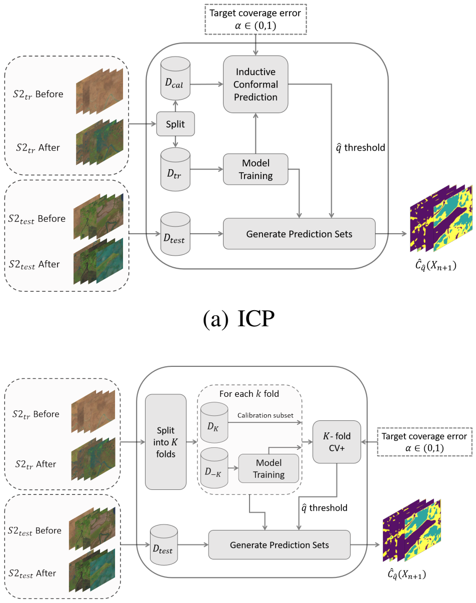
I. Introduction

Flood detection and delineation is essential for emergency response services, playing a fundamental role in the allocation of resources. Supervised machine learning methods, particularly deep neural networks like U-Net, excel in segmenting flood-affected areas from bitemporal satellite imagery [1, 2]. Despite their success in terms of accuracy however, an important question that remains open is how much can we trust the predictions of such models. In high-stakes scenarios like disaster response, the reliability of model predictions is as important as their accuracy since incorrect or overly confident predictions can lead to misallocation of resources or delayed actions, exacerbating the impact of the disaster.

Uncertainty quantification is vital for understanding and communicating model prediction reliability. To that end, approaches such as calibration methods [3, 4], Bayesian neural networks, and other probabilistic techniques [5] have been considered. Among these, Conformal Prediction (CP) has gained attention for its ability to provide statistically valid uncertainty estimates, offering guarantees by constructing **prediction sets** calibrated to a user-specified confidence level [6].

CP methods can be categorized based on computational and data requirements. Inductive Conformal Prediction (ICP) [6, 7] is a popular approach that achieves efficiency by splitting the data into training and calibration sets. However, this can degrade model performance in data-scarce situations. On the other hand, Full Conformal Prediction [6] avoids data splitting but is computationally expensive. To address this, methods like k-fold Cross-Validation Conformal Prediction (CV+) (closely related to Cross-Conformal Prediction [8]) and Jackknife techniques [9, 10] balance data usage with computational efficiency.

In the context of image segmentation, CP has been considered in medical imaging, e.g. for tumor segmentation [11] and tissue sub-region prediction [12]. Very recently, the concept of CP was introduced for quantifying uncertainty in land cover classification, canopy height estimation and invasive tree species mapping [13].



(a) ICP

(b) CV+

Fig. 1: Comparison of Two CP Methods: (a) Inductive CP (ICP), which splits the dataset into training and calibration sets, and (b) K-Fold Cross-Validation CP (CV+), which leverages cross-validation

This work was supported by the TITAN ERA Chair project (contract no. 101086741) within the Horizon Europe Framework Program of the European Commission.

In this work, we evaluated two CP methods in flood segmentation from Sentinel 2 observations, using the bitemporal U-Net model from Drakonakis et al. [2] as a baseline (see Figure 1). The key novelties of this work are as follows:

- Apply CP for uncertainty quantification in the analysis of remote sensing observations.
- Investigate different flavors of CP for bitemporal image segmentation.
- Experimentally demonstrate that different approaches to CP achieve different trade-offs relative to coverage, efficiency, and accuracy.

II. Conformal Prediction

A. Split Conformal Prediction

Split Conformal Prediction (Inductive CP) constructs prediction sets using a pre-trained model and a small calibration set. Given a model $\hat{f} : \mathcal{X} \to \mathbb{R}$ trained on (X, Y) and a calibration set $(X_1, Y_1), \ldots, (X_n, Y_n)$, for a new test sample X_{n+1}, the prediction set $C_\alpha(X_{n+1})$ is computed with a user-specified error rate α using non-conformity scores (e.g., $s_i = 1 - \hat{f}(X_i)_{Y_i}$). The prediction is included in the set if its score is below a threshold derived from the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the non-conformity scores. The marginal guarantee for the prediction set is:

$$P(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha$$

The only assumption required is that the calibration and test data together form an exchangeable sequence. This guarantee is *marginal*, meaning it holds on average over both the choice of the calibration dataset and the test sample.

A limitation is the potential loss of accuracy due to data splitting, as the model must remain independent of the holdout set.

B. K-Fold CV+ CP

We define the K-fold cross-validation method as presented in [10], in the context of regression. The training data is partitioned into K disjoint subsets of size $m = \frac{n}{K}$, and K models are constructed as follows:

$$\hat{\mu}_{-S_k} = \mathcal{A}\left((X_i, Y_i) \ : \ i \in \{1, \ldots, n\} \backslash S_k\right)$$

where $\mathcal{A}$ denotes the regression algorithm trained on all data except the k-th subset. The performance of the model is then evaluated using residuals derived from cross-validation:

$$R_i^{CV} = |Y_i - \hat{\mu}_{-S_{k(i)}}(X_i)|, \ i = 1 \ldots n$$

Here, $k(i)$ represents the index of the subset containing i. Based on these residuals, the prediction intervals are defined as:

$$\begin{aligned} \hat{C}^{CV+}_{n,K,\alpha}(X_{n+1}) = [&\hat{q}^-_{n,\alpha}\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) - R_i^{CV}\}, \\ &\hat{q}^+_{n,\alpha}\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{CV}\}] \end{aligned}$$

$\hat{q}^-_{n,\alpha}\{v_i\}$ and $\hat{q}^+_{n,\alpha}\{v_i\}$ are the $\lfloor (n+1)\alpha \rfloor / n$ and $\lceil (n+1)(1-\alpha) \rceil / n$ quantiles of the empirical distribution, respectively. The resulting K-fold CV+ prediction interval satisfies the following coverage guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha - \sqrt{2/n}$$

Notably, Jackknife+ can be regarded as a special case of the CV+ method with $K = n$.

III. Methods

A. Ombria-Net

We employed the Bitemporal OmbriaNet architecture for flood segmentation [2]. It extracts meaningful features from temporal changes in the input data, processing two images, one before and one after the event. We utilize the OMBRIA dataset, which consists of satellite imagery captured before and after a flood event. The dataset comprises imagery from the European Space Agency's (ESA) Copernicus program, offering global coverage, high spatial resolution, and frequent temporal updates [14]. For this experiment, we focus exclusively on applying CP to a model that processes images from Sentinel-2. The model is trained with binary cross-entropy loss, using the Adam optimizer with a learning rate of 0.0001 and batch size 8. With 10,796,485 parameters, Bitemporal OmbriaNet serves as a strong baseline for flood segmentation tasks.

B. Conformal Prediction Implementation for Flood Segmentation

In the context of image segmentation, the goal of conformal prediction is to construct prediction sets $\hat{C}_{\hat{q}}(X_{n+1})$ for a new image X_{n+1}, such that the true label for each pixel p is contained within the set with a desired confidence level $1 - \alpha$. To implement Conformal Prediction (CP) for flood segmentation, we evaluated the methods of Split Conformal Prediction (Split CP), and K-fold Cross-Validation Conformal Prediction (K-fold CV+). For Split CP, we split the dataset into training and calibration subsets with a proportion of 80%-20% between training and calibration. We use the non-conformity score

$$s_i = 1 - \hat{f}(X_i)_{Y_i}.$$

For K-fold CV+, we evaluated performance for $K = 5$, $K = 10$, and $K = 20$. To use this method in our segmentation setting, we follow the procedure of Cross-Validation Conformal Risk Control (CV-CRC) [15], but since we are only interested in conformal prediction and not general conformal risk control, we will use as loss l the miscoverage loss.

a) Threshold Computation: We start by estimating the population risk using cross-validation, with miscoverage loss for image j inside fold k

$$l(y_k[j], \hat{C}_{\hat{q}}(x_k[j]|D_{-k})) = \begin{cases} 1 & y_k[j] \notin \hat{C}_{\hat{q}}(x_k[j]|D_{-k}) \\ 0 & y_k[j] \in \hat{C}_{\hat{q}}(x_k[j]|D_{-k}) \end{cases}.$$

The threshold is then calculated as

$$\hat{q}_{n,\alpha} = \inf_q \left\{ q \middle| \hat{R}^{CV}(q|\mathcal{D}) \leq \alpha \right\}$$

where

$$\hat{R}^{CV}(q|\mathcal{D}) = \frac{1}{K+1}\sum_{k=1}^{K}\frac{K}{N}\sum_{j=1}^{N/K} l(y_k[j], \hat{C}_q^{CV}(x_k[j]|D_{-k})) + 1$$

b) Prediction Set Construction: Given a threshold $\hat{q}_{n,\alpha}$, the prediction set for an image $x_k[j]$ is calculated as

$$\hat{C}_{\hat{q}}^{CV}(x_k[j]|\mathcal{D}) = \{y' \in \mathcal{Y} : \min_k\{s^{CV}(x_k[j], y'|\mathcal{D}_{-k})\} \leq \hat{q}_{n,\alpha}\}$$

with

$$s^{CV}(x_k[j], y'|\mathcal{D}_{-k}) = 1 - \hat{f}_{\mathcal{D}_{-k}}(x_k[j])_{y'}.$$

Note that the guarantee derived from CV-CRC holds when $K \geq 1/\alpha - 1$, however we also experimentally use $K = 5$ for $\alpha = 0.1$, without losing coverage.

IV. Experiments

A. Evaluation Metrics

The evaluation focused on the empirical coverage achieved, the accuracy of the model, and the inefficiency, the size of the prediction set. We define empirical coverage and inefficiency, following the definition by [16], for a test set of size N:

$$\text{Cover} := \frac{1}{N}\sum_{i=1}^{N}\delta[y_i \in \mathcal{C}(x_i)], \quad \text{Inef} := \frac{1}{N}\sum_{i=1}^{N}|\mathcal{C}(x_i)|$$

where δ function is 1 if the argument is true, and 0 if it is false.

B. Naive vs CP

In this section, we highlight the benefits of using Conformal Prediction (CP) for constructing prediction sets, compared to relying on arbitrary threshold values. To quantify the benefits of CP, we generate diagrams showing the relationships between confidence and coverage, as well as confidence and inefficiency. These visualizations are inspired by the coverage diagrams presented in [17]. For the non-CP method, prediction sets are defined as:

$$\hat{C}_q(X_i) = \{y' \in \mathcal{Y} : \hat{f}(X_i, y') \geq q\}$$

where q is the chosen confidence threshold. In our experiments, we evaluate how coverage and inefficiency change as the confidence threshold varies. For instance, to achieve a coverage of 90%, the naive approach involves setting $q = 0.1$ and including all predictions above this threshold in the prediction set. Note that for $q = 0.5$, this reduces to plain segmentation, with coverage being equivalent to model accuracy. In contrast, CP directly ties the confidence level to the user-specified parameter, $(1 - \alpha)$.

Fig. 2 presents the coverage and inefficiency diagrams comparing the naive thresholding method and the K-Fold CV+ Conformal Prediction (CP) method (K=5). The results presented demonstrate that CP achieves coverage closer to the diagonal than the naive method. The naive method becomes overly conservative, resulting in coverage levels that exceed the desired values. This excessive conservatism is reflected in the inefficiency diagram, where inefficiency values are significantly higher, reaching as much as 2 when the coverage approaches 1. In contrast, CP demonstrates superior performance, aligning closely with the diagonal for confidence levels ≥ 0.8, while maintaining considerably lower inefficiency values.

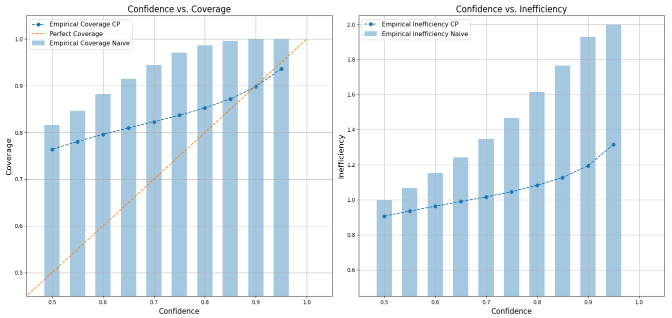


Fig. 2: Coverage Diagram for naive and K-Fold CV+ method, with $K = 5$

In practical scenarios, where higher coverage levels are typically preferred, the coverage deviation observed in the K-Fold method for lower confidence levels is less of a concern.

Figures 3, 4 provide a visual example of prediction sets formed using the K-Fold method ($K = 10$) for $\alpha = 0.1$ after 20 epochs of training, and prediction sets from arbitrary thresholds. In transition zones near flood boundaries, prediction sets frequently include both classes, reflecting the inherent ambiguity of these regions, while the sets shrink as thresholds approach 0.5.

The threshold acquired by CP in this case was close to 0.38. One would argue that instead of applying conformal prediction, we could simply create prediction sets for varying thresholds ≤ 0.5 and visually estimate the model's uncertain regions. However, without CP we would not know which threshold - prediction set - achieves the desired coverage. CP provides a distinct advantage by offering coverage guarantees, granting possible a rigorous uncertainty estimate. In cases where the model's prediction significantly deviates from the conformal prediction sets (as in figure 4), it may be wise to treat the prediction with caution.

C. ICP vs K-Fold CV+

We then compared the performance of Inductive Conformal Prediction (ICP) with K-Fold Cross-Validation Conformal Prediction (K-Fold CV+).As shown in Tables I and II, all conformal prediction (CP) methods achieve the expected coverage level at $\alpha = 0.1$. However, there are notable differences between ICP and K-Fold CV+ in terms of inefficiency and accuracy.

The ICP method exhibits higher inefficiency and lower accuracy compared to K-Fold CV+. This is primarily because ICP relies on data splitting, which reduces the size of the training set available for the model, leading to suboptimal performance. In contrast, K-Fold CV+ more efficiently utilizes the available data by partitioning it into folds. Each fold is used as a calibration set once, while the remaining data is

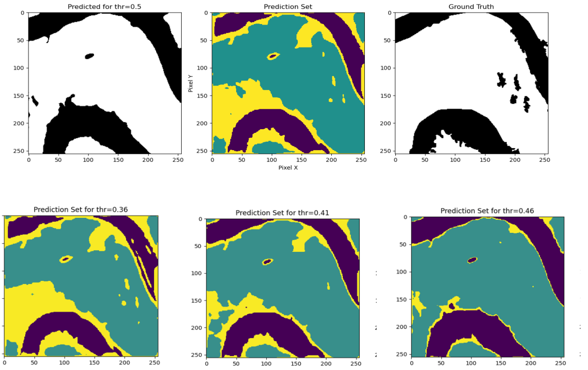

Predicted for thr=0.5
Prediction Set
Ground Truth
Pixel Y
Pixel X
Prediction Set for thr=0.36
Prediction Set for thr=0.41
Prediction Set for thr=0.46
0
50
100
150
200
250

Fig. 3: Example of prediction for K-Fold CV+ (top middle), $K = 10$, 20 epochs, and for arbitrary thresholds. Blue → only class "not flood" was included in the set, gray → only class "flood" was included in the set, yellow → both classes were included in the set. The K-Fold method captures uncertainty in ambiguous regions, such as transition zones near flood boundaries.

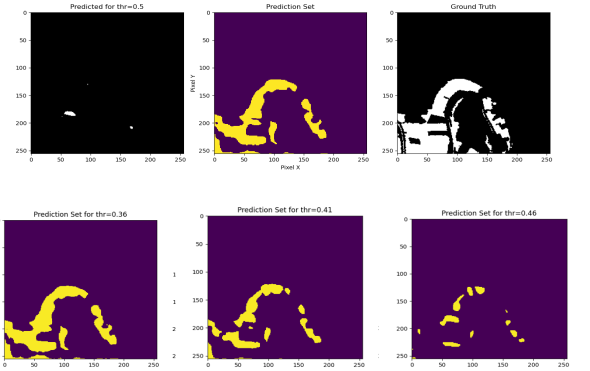

Predicted for thr=0.5
Prediction Set
Ground Truth
Pixel Y
Pixel X
Prediction Set for thr=0.36
Prediction Set for thr=0.41
Prediction Set for thr=0.46
0
50
100
150
200
250

Fig. 4: Example of prediction for K-Fold CV+ (top middle), $K = 10$, 20 epochs, and for arbitrary thresholds. Blue → only class "not flood" was included in the set, gray → only class "flood" was included in the set, yellow → both classes were included in the set.

used for training. This allows the model to be trained on more data in aggregate, which contributes to improved predictive performance and reduced inefficiency. Moreover, the results across different values of K (5, 10, and 20) show that the performance of K-Fold CV+ is relatively stable, indicating robustness to the choice of fold count.

Interestingly, the accuracy of the base OmbriaNet model is very close to the K-Fold CV+ results. This indicates that applying K-Fold CV+ does not degrade the underlying model's performance while adding the benefit of uncertainty quantification through valid prediction sets.

TABLE I: ICP vs K-Fold CV+ (5 Epochs)

Table	**OMBRIA Net for 5 epochs**		
CP-method, $\alpha = 0.1$	***Coverage***	***Inef.***	***Accuracy***
ICP (20% split train-cal)	0.9039	1.725	0.6663
K-Fold CP, K=5	0.9190	1.2832	0.8078
K-Fold CP, K=10	0.9044	1.2707	0.8021
K-Fold CP, K=20	0.9043	1.2589	0.8037
OmbriaNet Base	–	–	0.8012

TABLE II: ICP vs K-Fold CV+ (20 Epochs)

Table	**OMBRIA Net for 20 epochs**		
CP-method, $\alpha = 0.1$	***Coverage***	***Inef.***	***Accuracy***
ICP (20% split train-cal)	0.8989	1.664	0.7021
K-Fold CP, K=5	0.8980	1.1935	0.8153
K-Fold CP, K=10	0.8887	1.1279	0.8196
K-Fold CP, K=20	0.9022	1.1993	0.8157
OmbriaNet Base	–	–	0.8164

V. Conclusions

In this work, we explore the advantages of using Conformal Prediction (CP) to construct prediction sets for uncertainty quantification, and how K-Fold CV+ compares to the more common ICP. Our experiments demonstrate that K-Fold CV+ methods, which effectively use cross-validation, perform well in providing accurate and efficient prediction sets, while ICP methods, although reliable in coverage and simply implemented, exhibit higher inefficiency and lower accuracy due to data splitting.

Although K-Fold CV+ is more computationally demanding than ICP, its benefits in accuracy and uncertainty estimation make it a worthwhile trade-off in many practical scenarios. Overall, our results suggest that in cases with limited data, as in our experiments, K-Fold CV+ is the preferred approach, providing more reliable and accurate prediction sets.

It is worth mentioning that the proposed framework has broader applicability beyond flood detection tasks, such as land cover classification, vegetation monitoring, and urban mapping. By integrating observations from multiple instruments, such as radar and optical sensors, CP could enhance model robustness and uncertainty quantification.

References

[1] G. Konapala, S. V. Kumar, and S. K. Ahmad, "Exploring sentinel-1 and sentinel-2 diversity for flood inundation mapping using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 163–173, 2021.

[2] G. I. Drakonakis, G. Tsagkatakis, K. Fotiadou, and P. Tsakalides, "Ombrianet—supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2341–2356, 2022.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[4] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration

of modern neural networks," *Advances in neural information processing systems*, vol. 34, pp. 15 682–15 694, 2021.

[5] T. J. Sullivan, *Distributional Uncertainty*. Cham: Springer International Publishing, 2015, pp. 295–318. [Online]. Available: https://doi.org/10.1007/978-3-319-23395-6_14

[6] *Conformal prediction*. Boston, MA: Springer US, 2005, pp. 17–51. [Online]. Available: https://doi.org/10.1007/0-387-25061-1_2

[7] H. Papadopoulos, "Inductive conformal prediction: Theory and application to neural networks," in *Tools in artificial intelligence*. Citeseer, 2008.

[8] V. Vovk, "Cross-conformal predictors," *Annals of Mathematics and Artificial Intelligence*, vol. 74, pp. 9–28, 2015.

[9] Y. Romano, M. Sesia, and E. Candes, "Classification with valid and adaptive coverage," *Advances in neural information processing systems*, vol. 33, pp. 3581–3591, 2020.

[10] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021.

[11] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=33XGfHLtZg

[12] H. Wieslander, P. J. Harrison, G. Skogberg, S. Jackson, M. Fridén, J. Karlsson, O. Spjuth, and C. Wählby, "Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 371–380, 2021.

[13] 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," *Scientific Reports*, vol. 14, no. 1, p. 16166, 2024.

[14] M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen, "Esa's sentinel missions in support of earth system science," *Remote Sensing of Environment*, vol. 120, pp. 84–90, 2012, the Sentinel Missions - New Opportunities for Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003442571200065X

[15] K. M. Cohen, S. Park, O. Simeone, and S. S. Shitz, "Cross-validation conformal risk control," in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 250–255.

[16] D. Stutz, K. D. Dvijotham, A. T. Cemgil, and A. Doucet, "Learning optimal conformal classifiers," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=t8O-4LKFVx

[17] J. Brunekreef, E. Marcus, R. Sheombarsing, J.-J. Sonke, and J. Teuwen, "Kandinsky conformal prediction: efficient calibration of image segmentation algorithms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4135–4143.