Beyond Morphological Dilation: Revisiting Conformal Prediction for 3D Tumor Segmentation

Luisa Vargas[1 [0009−0006−5741−8485]], Simone Rossi[1 [0000−0003−2908−3703]], and Maria A. Zuluaga[1 [0000−0002−1147−766X]]

EURECOM, Biot, France {luisa.vargas, simone.rossi, maria.zuluaga}@eurecom.fr

Abstract. Deep learning has achieved remarkable performance for tumor segmentation, yet high overlap scores do not necessarily imply reliable boundary delineation or complete identification of disconnected tumor foci. While uncertainty quantification methods provide voxel-wise confidence estimates, conformal prediction (CP) offers finite-sample coverage guarantees through calibrated prediction sets. Existing CP methods for image segmentation commonly rely on morphological dilation. We show that this construction becomes *vacuous* in the presence of disconnected segmentation errors: missed tumor components saturate the calibrated score, producing prediction sets that preserve formal coverage while becoming operationally uninformative. We characterize this failure mode through the distinction between formal and operational coverage, and propose two complementary extensions. Instance-level CP separates boundary and structural errors, whereas Geodesic CP replaces isotropic dilation with confidence-guided expansion, and we combine both in a unified formulation. On BraTS 2021, across 100 calibration/test splits, the unified formulation eliminates vacuity while preserving split-conformal validity. The code is available at: https://github.com/robustml-eurecom/conformal-tumor-segmentation.git

Keywords: Conformal prediction · Tumor segmentation · Geodesic distance · Boundary uncertainty

1 Introduction

Deep learning has achieved remarkable performance for the segmentation of compact anatomical structures [7] and tumors [4, 10], with benchmark scores often approaching expert-level agreement. Yet high accuracy does not necessarily imply reliable boundary delineation. Standard overlap metrics are driven by the correctly segmented interior of the object, while localized boundary errors or small disconnected tumor foci may have only a limited effect on the reported performance. In clinical applications such as surgical planning and radiotherapy, however, accurate delineation of the tumor boundary is essential, as segmentation errors can propagate to treatment margins and affect downstream plan-

ning [13, 15]. Equally important, failing to identify a small disconnected tumor focus may result in an incomplete delineation of the tumor extent.

Uncertainty quantification (UQ) methods have been widely adopted to estimate the reliability of segmentation predictions. Most approaches generate voxel-wise uncertainty maps, estimating the confidence of the model's predictions at each voxel. Such maps are particularly useful for highlighting uncertain boundaries, where segmentation errors are most likely to occur. Bayesian approximations and deep ensembles are among the most common approaches for obtaining these uncertainty estimates [12]. While informative, these methods provide confidence scores rather than finite-sample guarantees on the segmented object.

Conformal prediction (CP) offers an alternative perspective. Rather than assigning confidence to individual voxels, CP calibrates a non-conformity score on held-out data and returns a prediction set with marginal coverage at least $1 - \alpha$, under exchangeability and for any fixed model [2, 16]. This guarantee is model-agnostic and distribution-free. However, the usefulness of the resulting prediction set depends on the choice of non-conformity score.

For image segmentation, prior CP methods typically construct prediction sets by calibrating either a dilation radius, applied through morphological dilation of the predicted mask [11], or a threshold on voxel-wise confidence maps [1, 3, 6, 8]. Morphological dilation uniformly expands the predicted mask in all directions, irrespective of the local confidence. Furthermore, when a disconnected tumor focus is missed during calibration, no finite local dilation can recover it, causing the corresponding calibration score to saturate. As a result, the calibrated radius may be driven by these pathological cases rather than by the boundary uncertainty it is intended to quantify.

Motivated by these limitations, we revisit conformal prediction for image segmentation. We show that morphological dilation can become *vacuous* in the presence of disconnected segmentation errors, producing prediction sets that preserve formal coverage guarantees while becoming operationally uninformative. We therefore propose two complementary extensions: **Instance-level CP**, which disentangles boundary uncertainty from missed disconnected components, and **Geodesic CP**, which replaces isotropic dilation with confidence-guided expansion. Finally, we combine both ideas in a unified formulation.

2 Methods

2.1 Split Conformal Prediction for Segmentation

Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ be an exchangeable calibration set and let $s(X, Y)$ be a non-conformity score, where larger values indicate poorer agreement between a prediction and its ground truth [16, 2]. Split conformal prediction computes the scores $s_i = s(X_i, Y_i)$ and the finite-sample quantile

$$\hat{q} = \text{the } \lceil (n+1)(1-\alpha) \rceil\text{-th smallest value of } \{s(X_i, Y_i)\}_{i=1}^n, \tag{1}$$

with $\hat{q} = \infty$ if the requested order is larger than n, and returns $C(X) = \{y : s(X, y) \leq \hat{q}\}$. Then $P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$, providing a finite-sample marginal coverage guarantee [2, 16].

For image segmentation, the prediction set is obtained by calibrating a nested family of candidate segmentations $\{C_\lambda(\hat{Y})\}_\lambda$, where λ controls the amount of expansion around the predicted mask $\hat{Y}$. The corresponding non-conformity score is defined as

$$s(X, Y) = \inf \left\{ \lambda : \frac{|Y \cap C_\lambda(\hat{Y})|}{|Y|} \geq \tau \right\}, \tag{2}$$

where $\tau = 1$ means full containment.

Different conformal segmentation methods differ only in the construction of the nested family $C_\lambda(\hat{Y})$. Morphological CP uses $C_\lambda(\hat{Y}) = \mathrm{dil}(\hat{Y}, \lambda)$ [11]. Threshold-based CP uses confidence sets such as $C_\lambda(\hat{Y}) = \{x : \hat{p}(x) \geq 1 - \lambda\}$, calibrated through risk control [3, 6, 8]. Throughout this work, the split-conformal calibration procedure remains unchanged; our contributions lie in the design and analysis of alternative non-conformity scores.

2.2 Morphological Conformal Prediction and Vacuity

Morphological CP constructs prediction sets by calibrating the smallest dilation radius that covers a fraction τ of the tumor:

$$s_{\mathrm{dil}} = \inf \left\{ r \in \mathbb{N}_0 : \frac{|Y_{\mathcal{R}} \cap \mathrm{dil}(\hat{Y}_{\mathcal{R}}, r)|}{|Y_{\mathcal{R}}|} \geq \tau \right\}. \tag{3}$$

In practice, s_{dil} is evaluated on a finite radius grid up to $r_{\max}$ (we use $r_{\max} = 20$ voxels). When the model misses a disconnected focus, no radius up to $r_{\max}$ reaches coverage τ: the smallest covering radius does not exist (formally the infimum is $+\infty$), so the score is set to the cap $r_{\max}$. The cap is therefore needed to define the score on these cases, not an optional regularizer, and it preserves exchangeability because s_{dil} remains a deterministic function of each fixed prediction and its ground truth.

Although score capping preserves the formal split-conformal guarantee, disconnected segmentation errors can cause the calibrated score to no longer exclusively reflect boundary uncertainty. To characterize the resulting behavior, we distinguish between formal and operational coverage. Formal coverage is the CP event $s \leq \hat{q}$; operational coverage asks whether the deployed set actually contains the target fraction of the tumor. The two agree when the score is finite. They disagree when more than α of calibration cases hit the cap: the quantile then sits at $r_{\max}$, so CP still covers the capped score while the deployed set need not cover the tumor. We refer to this divergence between formal and operational coverage as *vacuity*. Formal coverage certifies the capped score $\min(s_{\mathrm{dil}}, r_{\max})$, not the original containment event; once the calibrated quantile reaches $r_{\max}$ the two events no longer coincide.

A capped dilation score mixes two different events: a boundary error that can be corrected by adding a margin, and a topological or detection error in

which an entire focus is disconnected from the prediction. In the second case, increasing the local margin around $\hat{Y}_{\mathcal{R}}$ may add many background voxels without reaching the missed focus. We therefore treat saturation as a diagnostic of score misspecification, rather than as a failure of CP itself.

2.3 Instance-level Conformal Prediction

Instance-level CP addresses the vacuity of morphological CP by separating boundary errors from disconnected missed components and calibrating them independently through spatial and structural scores. Given the connected components of $Y_{\mathcal{R}}$, a component is *matched* if it overlaps $\hat{Y}_{\mathcal{R}}$ by at least κ voxels (default $\kappa = 1$), and *unmatched* otherwise; Sec. 4 reports robustness to κ. Let M be the union of matched components and $U = Y_{\mathcal{R}} \setminus M$. We define

$$s_{\text{spatial}} = \inf \left\{ r : \frac{|M \cap \text{dil}(\hat{Y}_{\mathcal{R}}, r)|}{|M|} \geq \tau \right\}, \tag{4}$$

$$s_{\text{struct}} = \frac{|U|}{|Y_{\mathcal{R}}|} \in [0, 1]. \tag{5}$$

The spatial score ignores unreachable missed foci, while the structural score records how much tumor was missed. We calibrate both scores at level $\alpha/2$, giving r^* and β^*. Cases with no matched component satisfy the spatial clause by convention, and β^* is calibrated over all cases, so a union bound certifies the marginal joint event

$$P\left(\frac{|M \cap \text{dil}(\hat{Y}_{\mathcal{R}}, r^*)|}{|M|} \geq \tau \text{ and } s_{\text{struct}} \leq \beta^* \right) \geq 1 - \alpha. \tag{6}$$

This is a two-part certificate, not a full-containment claim $Y \subseteq C$: detected components are covered at τ, and the missed-volume fraction is bounded by β^*. The equal $\alpha/2 + \alpha/2$ split is the Bonferroni allocation; it is assumption-free but conservative when the two clauses tend to hold together. Any fixed split summing to α is equally valid, and a tighter data-driven allocation would require a separate calibration split, since choosing the split on the same data used for r^* and β^* would void the guarantee.

2.4 Geodesic Conformal Prediction

Geodesic CP replaces isotropic dilation with a confidence-guided geodesic distance, allowing prediction sets to expand preferentially along paths of higher predicted confidence. The geodesic distance is represented by an arrival-time function $T(x)$, obtained by solving the Eikonal equation $\|\nabla T(x)\| = 1/F(x)$, where the propagation speed is defined as

$$F(x) = \hat{p}_{\mathcal{R}}(x) + \varepsilon, \qquad \varepsilon = 0.05, \tag{7}$$

with $T = 0$ on $\hat{Y}_{\mathcal{R}}$. The Eikonal equation is solved using the Fast Marching algorithm [14]. The floor $\varepsilon > 0$ keeps the speed positive everywhere, so the front crosses low-probability background and reaches disconnected foci in finite time; without it a missed component would receive an unbounded score. As ε grows, F approaches a constant and the geodesic set approaches isotropic dilation, so ε sets the contrast between confident tumor ($F \approx 1 + \varepsilon$) and background ($F \approx \varepsilon$), about 21 at $\varepsilon = 0.05$.

The score is the τ-quantile of T over $Y_{\mathcal{R}}$ and the set is $\{x : T(x) \leq \lambda^*\} \cup \hat{Y}_{\mathcal{R}}$.

By construction, the geodesic score remains finite for disconnected missed components, eliminating the need for score capping. Furthermore, the confidence-guided geometry allows the calibrated prediction set to expand anisotropically, following regions of high predicted confidence instead of applying a uniform Euclidean margin.

2.5 Unified Formulation

We combine the instance decomposition with confidence-guided spatial calibration by replacing the matched-component dilation score with the matched-component geodesic score, while leaving the structural score unchanged. The resulting formulation combines confidence-guided boundary expansion with explicit quantification of disconnected missed tumor volume.

3 Experimental Setup

Data and protocol. We use the BraTS 2021 training set [10, 4, 5], with four MRI channels and labels $\{0, 1, 2, 4\}$. A transformer segmentation model trained on BraTS 2021 is used to generate discrete segmentation masks together with softmax probability maps for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET). These predictions are generated once and kept fixed throughout all experiments [9].

From the 1,251 BraTS 2021 training cases, we retain the 1,239 for which the fixed model produced a complete hard prediction and softmax map; the remaining 12 lacked one of these outputs and were excluded. Each split uses 991 cases for calibration and 248 for testing. We set $\alpha = 0.1$ and report means over 100 random 80/20 splits. Since the segmentation model is fixed before calibration, the calibration/test split is the only source of randomness in the conformal guarantee.

Ground-truth components. Within each retained case, ground-truth components below 100 voxels (0.1 ml at $1\,\mathrm{mm}^3$) are removed before scoring, since at this scale they are not reliably separable from BraTS annotation noise. Retaining sub-threshold foci would only add structural misses, so the reported gains are a lower bound.

We compare two existing conformal approaches based on morphological dilation and confidence thresholding with the proposed Instance-level CP, Geodesic CP, and unified formulation. CP is evaluated separately for WT, TC, and ET.

Table 1: Full-containment dilation CP at $\tau = 1$ (100 splits, target coverage 0.90). Regions: WT (whole tumor), TC (tumor core), ET (enhancing tumor). *Capped frac.*: fraction of cases whose dilation score reaches the cap $r_{\max} = 20$. *Formal cov.*: coverage of the (capped) conformal score, $s \leq \hat{q}$. *Operational cov.*: fraction of test cases whose deployed set contains the full tumor. *Vacuous splits*: splits whose calibrated radius saturates while operational coverage stays below target.

Region	Capped frac.	Formal cov.	Operational cov.	Vacuous splits
WT	0.157	1.000	0.842	100%
TC	0.091	0.906	0.905	2%
ET	0.098	0.935	0.899	43%

All methods use the same predictions and softmax maps; only the non-conformity score differs. Connected components are computed using 26-connectivity, whereas morphological dilation uses a 6-connected neighborhood. The dilation grid is capped at $r_{\max} = 20$ and the geodesic speed floor is $\varepsilon = 0.05$. We report formal coverage, operational coverage, prediction-set volume in voxels, and the fraction of vacuous splits. All reported volumes include the original predicted mask plus the conformal margin. For instance-level methods, the spatial volume corresponds to the returned set around the prediction; the structural bound β^* is reported through the certificate rather than added as a separate spatial region, because unmatched components have no observed location at test time.

4 Results

Full-containment dilation can be vacuous. Table 1 shows dilation CP at $\tau = 1$. For WT, 15.7% of cases hit the cap, above $\alpha = 0.1$. The calibrated radius therefore saturates in every split: formal coverage is 1.000, but operational coverage is only 0.842. ET sits at the boundary: the capped fraction is 9.8%, just below α, so split-to-split fluctuation makes 43% of splits vacuous; the same instability appears in the formal coverage, which varies across splits (0.935 ± 0.057) as the quantile crosses the cap. TC is effectively non-vacuous. The difference between regions is consistent with their anatomy and with the segmentation task. WT is larger and more heterogeneous, so a small disconnected residual can be far from the predicted mask. ET is often small and fragmented, so missing one focus can strongly affect full containment even when the absolute missed volume is limited. TC lies between these regimes.

The cap does not create the vacuity. For WT the capped fraction (0.157) exceeds α, while a component is fully missed in only 0.095 of cases, so saturation reflects both missed foci and the boundary tail of matched components. Raising $r_{\max}$ lowers the capped fraction but not the vacuity, since reaching a disconnected focus requires dilating through background; the geodesic score avoids this by remaining finite without a cap.

Table 2: Operational coverage (Cov.), prediction-set volume in voxels (Vol.), and fraction of vacuous splits (Vac.) at $\tau = 1$ and $\tau = 0.95$, over 100 splits. Cov. at $\tau = 1$ is $\text{mean}_{\pm\text{sd}}$. For Instance and Unified, Cov. is the joint guarantee of Eq. (6); for Threshold, Dilation, and Geodesic it is operational coverage at the stated τ. Bold marks the smallest-volume non-vacuous method. † marks methods vacuous in the majority of splits.

		$\tau = 1$			$\tau = 0.95$	
Region	Method	Cov.	Vol.	Vac.	Cov.	Vol.
WT	Threshold	$.900_{\pm.022}$	8910	0%	.901	2489
	Dilation	$.842^{\dagger}_{\pm.022}$	426	100%	.914	135
	Geodesic	$.900_{\pm.023}$	**497**	0%	.900	**118**
	Instance	$.853^{\dagger}_{\pm.021}$	426	100%	.919	160
	Unified	$.911_{\pm.019}$	524	0%	.913	138
TC	Threshold	$.901_{\pm.021}$	8921	0%	.904	6713
	Dilation	$.905_{\pm.021}$	205	2%	.909	47
	Geodesic	$.901_{\pm.023}$	163	0%	.900	**43**
	Instance	$.910_{\pm.018}$	174	0%	.908	71
	Unified	$.909_{\pm.019}$	**125**	0%	.906	47
ET	Threshold	$.905_{\pm.024}$	8928	0%	.900	8608
	Dilation	$.899^{\dagger}_{\pm.019}$	192	43%	.912	37
	Geodesic	$.899_{\pm.022}$	161	0%	.900	**31**
	Instance	$.911_{\pm.021}$	132	0%	.921	44
	Unified	$.907_{\pm.020}$	**106**	0%	.909	35

Main comparison. Table 2 compares all methods. At $\tau = 1$, thresholding reaches the target but returns near-whole-image sets (about 8,910 voxels on WT), so its coverage is not operationally informative. Dilation is vacuous on WT in every split and on 43% of ET splits. The two proposed methods address different failure modes. Instance-level CP targets structural misses, so it removes the ET vacuity and reduces the ET set from 192 to 132 voxels; the unified method reduces it further to 106 voxels, 45% below dilation. Instance-level CP does not, however, recover WT (coverage 0.853, still vacuous), because WT saturation comes from the boundary-distance tail of *matched* components rather than from missed components. Geodesic CP targets that tail and is the smallest valid method on WT, while the unified method is the smallest on TC and ET. On WT, valid full containment costs volume: the non-vacuous geodesic (497 voxels) and unified (524 voxels) sets exceed the vacuous dilation set (426 voxels) because they cover the tail instead of truncating it at the cap. At $\tau = 0.95$ every method is non-vacuous, and the spatial methods are far smaller than thresholding, with geodesic being the tightest in all three regions.

Separated guarantees. Instance-level and unified sets provide a two-part guarantee: coverage of the matched components at τ and a bound β^* on the fraction

Table 3: Instance-level and unified guarantees at $\tau = 1$ over 100 splits, reported as separate events. Spatial cov.: coverage of matched components at τ (target 0.95). Struct. cov.: cases with missed fraction within β^* (target 0.95). Joint cov.: both clauses hold (target 0.90). r^*: calibrated matched-component radius (instance only; cap 21). β^*: calibrated bound on the missed-volume fraction, in percent.

Region	Method	Spatial cov.	Struct. cov.	Joint cov.	r^*	β^* (%)
WT	Instance	.883	.952	.853	21.0	0.64
	Unified	.949	.952	.911	–	0.64
TC	Instance	.955	.951	.910	17.1	0.43
	Unified	.952	.951	.909	–	0.43
ET	Instance	.954	.950	.911	14.1	1.72
	Unified	.949	.950	.907	–	1.31

Table 4: Instance-level CP sensitivity to the matching threshold κ (minimum predicted–GT overlap in voxels to call a component matched) at $\tau = 1$, over 100 splits. Joint coverage and volume are stable; stricter matching moves boundary-touching foci into the structural term, raising β^* and the fraction of cases with a missed component. Volume in voxels.

Region	κ	Joint cov.	Vol.	β^* (%)	Missed-case frac.
WT	1	.853	426	0.64	.095
	5	.853	426	0.66	.096
	20	.852	426	0.74	.101
TC	1	.910	174	0.43	.323
	5	.910	174	0.50	.340
	20	.911	174	0.57	.356
ET	1	.911	132	1.72	.461
	5	.911	132	1.93	.527
	20	.911	132	2.41	.546

of tumor missed entirely. Table 3 reports both clauses and their joint event, each near its target, except the instance-level WT spatial clause (0.883), where the matched components saturate at the cap ($r^* = 21$); the geodesic margin in the unified method recovers WT (0.949 spatial, 0.911 joint). The bound β^* stays small (0.4 to 1.7% of tumor volume) although misses are frequent, so it reports the missed fraction explicitly rather than inflating the radius.

Matching threshold. Table 4 varies the overlap κ required to call a component matched. Joint coverage varies by at most 0.001 and the set size is unchanged across $\kappa \in \{1, 5, 20\}$; increasing κ only moves boundary-touching foci into the structural clause, raising β^* without changing the joint event.

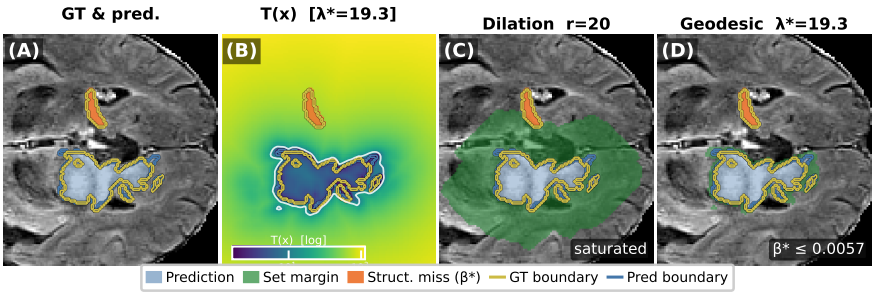

Fig. 1: WT multi-focal example at $\tau = 0.95$. (A) GT and prediction, missed focus in orange. (B) arrival-time field $T(x)$ with the λ^* level set. (C) dilation at the case radius, which misses the focus. (D) geodesic set at λ^*; the missed fraction is certified by β^*.

Qualitative interpretation. Figure 1 shows the construction at $\tau = 0.95$ on a multi-focal WT case. Dilation forms an isotropic band whose radius is set by the farthest ground-truth voxel; when that voxel belongs to a disconnected missed focus, the required radius exceeds the grid cap, so the focus stays outside the deployed set. Geodesic CP still expands from the predicted mask, but its arrival-time field uses the softmax to prefer plausible tumor-adjacent paths. The unified method then avoids conflating two errors: the visible margin describes spatial uncertainty around the detected component, while β^* bounds the fraction of missed tumor volume. Boundary uncertainty and missed-component uncertainty are different error types, and the set should communicate both rather than collapse them into one radius when the score cannot represent both faithfully.

Reliability and structural misses. Figure 2 illustrates two main findings. Per-split operational coverage fluctuates by about 0.02, as expected for this calibration size; on WT, dilation and instance-level CP stay below the target while the unified method reaches it. The structural score shows that fully missed components are common, especially for ET: at least one focus is missed in 10% of WT, 32% of TC, and 46% of ET cases. This confirms that missed disconnected components are a model failure, not only a calibration artifact. These per-case miss rates summarize the model's lesion-level detection. The instance-level and unified certificates make the failure visible through β^* rather than hiding it via an inflated radius.

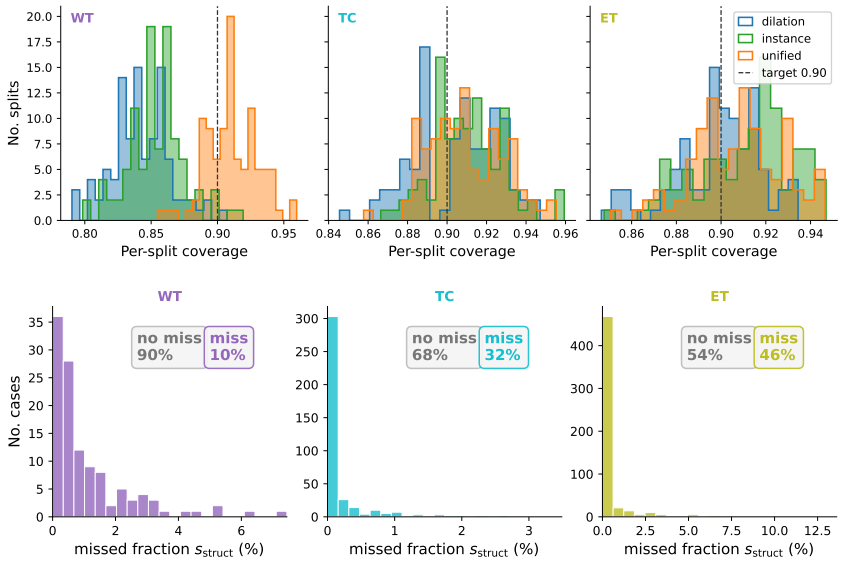

Fig. 2: (top) Per-split operational coverage for dilation, instance-level CP, and unified, per region; dashed line is target 0.90. WT: dilation and instance-level CP fall below target, the unified method reaches it. (bottom) Per-case distribution of $s_{\text{struct}} = |U|/|Y_{\mathcal{R}}|$, with fraction of cases with no missed component and with at least one.

5 Conclusion

We identify a failure mode of morphological conformal prediction, which we term *vacuity*: prediction sets may preserve formal coverage while becoming operationally uninformative in the presence of disconnected segmentation errors. We characterized this failure mode through the distinction between formal and operational coverage, and proposed two complementary extensions. Instance-level CP separates boundary and structural errors, Geodesic CP replaces isotropic dilation with confidence-guided expansion, and their unified formulation combines both ideas into a single conformal framework, eliminating vacuity across all tumor regions on BraTS 2021 while preserving split-conformal validity.

Our guarantees remain marginal and rely on exchangeability between calibration and test data, requiring recalibration under distribution shift. More broadly, our results suggest that conformal segmentation methods should be evaluated not only by coverage and prediction-set size, but also by whether the resulting prediction sets remain operationally meaningful. While our study focuses on BraTS 2021, assessing the prevalence of vacuity across other segmentation tasks and anatomical structures is an important direction for future work.

References

1. Angelopoulos, A., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: International conference on learning representations. vol. 2024, pp. 55198–55218 (2024)
2. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023). https://doi.org/10.1561/2200000101
3. Angelopoulos, A.N., Kohli, A.P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., Romano, Y.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning. pp. 717–730. PMLR (2022)
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
5. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
6. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.I.: Distribution-free, risk-controlling prediction sets. Journal of the ACM **68**(6) (2021). https://doi.org/10.1145/3478535
7. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
8. Kutiel, G., Cohen, R., Elad, M., Freedman, D., Rivlin, E.: Conformal prediction masks: Visualizing uncertainty in medical imaging. In: International Conference on Learning Representations (ICLR) (2023)
9. Luu, H.M., Park, S.H.: Extending nn-unet for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 173–186. Springer International Publishing, Cham (2022)
10. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
11. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. arXiv preprint arXiv:2503.05618 (2025)
12. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Analysis **59**, 101557 (2020). https://doi.org/10.1016/j.media.2019.101557
13. Njeh, C.F.: Tumor delineation: The weakest link in the search for accuracy in radiotherapy. Journal of Medical Physics **33**(4), 136–140 (2008). https://doi.org/10.4103/0971-6203.44472
14. Sethian, J.A.: Fast marching methods. SIAM Review **41**(2), 199–235 (1999). https://doi.org/10.1137/S0036144598347059
15. Vinod, S.K., Jameson, M.G., Min, M., Holloway, L.C.: Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiotherapy and Oncology **121**(2), 169–179 (2016). https://doi.org/10.1016/j.radonc.2016.09.009

16. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)