Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trustworthy Vertebral Segmentation in Spine CT via External Validation and Per-Vertebra Reliability Assessment

Mustafa Farooqui

fau_mfarooqui@fau.edu

Florida Atlantic University

Maohua Lin

Florida Atlantic University

Frank Vrionis

Florida Atlantic University

Yufei Tang

Florida Atlantic University

Min Shi

Florida Atlantic University

Research Article

Keywords: spine CT, vertebral instance segmentation, nnU-Net, external validation, ensemble uncertainty, quality control

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10639333/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose. Clinical adoption of automated vertebral segmentation from spine CT depends on two forms of trust: performance under genuine external validation rather than same-source testing, and a signal for when an individual prediction is unreliable. We evaluate nnU-Net, decompose the cross-domain gain into training-data expansion versus ensembling, and test whether per-vertebra ensemble disagreement can flag vertebrae for review.

Methods. An nnU-Net v2 residual-encoder (ResEnc-M) model was trained on VerSe alone (112 cases) and on an expanded VerSe-plus-CTSpine1K set (387 cases), then evaluated on the held-out CTSpine1K validation split (191 cases), external to the VerSe-only model. Single-model and five-fold ensemble inference were compared. Per-vertebra Dice (DSC) was computed against reference labels; ensemble disagreement was tested against DSC by Spearman correlation and AUROC with bootstrap CIs.

Results. The expanded ensemble reached a mean per-vertebra Dice of 0.9206 (lumbar 0.957, cervical 0.888). Five-fold ensembling raised single-source Dice from 0.8315 to 0.8779 (+ 4.6), exceeding the gain from adding a second source (0.8578, + 2.6). Ensemble disagreement predicted per-vertebra error (Spearman 0.862; AUROC 0.920); flagging vertebrae below 0.8 fold agreement captured 76.1% of poorly segmented vertebrae while referring 9.4% for review.

Conclusion. Ensemble disagreement provides a validated, cost-free per-vertebra reliability flag that predicts segmentation error and localizes review, catching about three-quarters of poor vertebrae while referring under one-tenth. Ensembling drove more cross-domain gain than added data here. The emphasis is trustworthy deployment of an existing model, not a new one.

1. Introduction

Automated vertebral instance segmentation from spine CT underpins quantitative spinal analysis, supporting fracture assessment [1], morphometry, surgical planning, and opportunistic screening. The self-configuring nnU-Net framework [2] is the de facto standard, with challenge-winning VerSe methods reporting mean Dice similarity coefficients (DSC) near 0.90 across cervical, thoracic, and lumbar levels [3–5].

Three gaps limit clinical trust. First, performance is usually measured on in-distribution test data from the same source as training [6, 7], and multi-site studies show models degrade under cross-site distribution shift [8]; recent benchmarking argues validation quality should weigh as heavily as architectural novelty, and that fair comparison requires isolating an innovation from confounders such as added data or altered inference [9], yet genuine external validation remains underused for whole-spine CT. Second, when a model gains from both more training data and ensembling, the combined improvement hides each factor's contribution, though the two differ sharply in cost. Third, deployment requires a model to signal when its output is unreliable, so questionable predictions can be routed for review rather than propagated into fracture reads, morphometry, or planning.

Recent spine work has begun to address reliability: structural uncertainty flags unreliable vertebrae in myeloma CT [10], and vertebra-aware frameworks add quality-control-aware outputs [11]. Cross-validation fold disagreement has served as a training-free quality signal. Gottlich et al. [12] collapsed interfold Dice into a per-image score to flag poor scans in kidney and tumor segmentation, and related work established ensemble disagreement as a failure detector under distribution shift [13–15], though it conflates model uncertainty with data-subsampling variability [16].

We do not propose a new segmentation architecture; our contribution is a framework for trustworthy deployment of an existing one, centered on a per-vertebra reliability flag. A standard nnU-Net run already trains five cross-validation folds, and where those folds disagree on a given vertebra, its segmentation is more likely to be wrong. We test whether that disagreement predicts segmentation error, quantify its value as a human-in-the-loop screening signal, and characterize when it fails. Supporting this, we evaluate all configurations under genuine external validation on the CTSpine1K public validation split, use a controlled four-condition ablation to separate training-data expansion from five-fold ensembling, and report both all-label and presence-matched metrics to separate true error from annotation-boundary artifacts.

2. Materials and Methods

2.1 Datasets

Two public spine CT datasets were used. VerSe 2019/2020 [3–5] comprises 374 multi-detector CT scans from 355 patients on scanners from four manufacturers, with voxel-level vertebra labels; the VerSe-only model used the standard 112-case training split, and the expanded set used all 141 annotated cases. CTSpine1K [17] is a multi-source collection of 1,005 CT volumes drawn from CT Colonography [18], HNSCC-3DCT-RT [19], Medical Segmentation Decathlon task03 liver [20], and a COVID-19 chest CT cohort [21]. Its 197-case public validation split was held out exclusively for external evaluation; the 246 CTSpine1K training cases came from the disjoint 610-case training split, yielding a genuine cross-domain evaluation for the VerSe-only model.

The expanded training set (Dataset002_SpineExpanded) combined 141 VerSe with 246 CTSpine1K cases (387 total). During curation, boundary-sliver labels (vertebra labels clipped at the field-of-view edge and retaining below 100 voxels) were zeroed in the training ground truth by a seeded script: 38 such labels across 31 cases were set to background, while all other labels and full cases were retained, and no complete case was removed. Initial development also used an internal VerSe held-out set of 28 cases, but all reported metrics derive from the external CTSpine1K split.

2.2 Segmentation model and training

Segmentation used nnU-Net v2 [2] in the 3d_fullres configuration with the residual-encoder [22] ResEnc-M plan (nnUNetResEncUNetMPlans), a U-Net-family architecture [23] released as a standardized baseline in the nnU-Net revisited benchmark [9]. nnU-Net self-configures preprocessing, patch and batch

size, and schedule from dataset fingerprinting; no manual tuning was applied. Training used a compound Dice plus cross-entropy loss under two regimes: a single model on all cases (fold_all), and a five-fold cross-validation ensemble combined by per-voxel majority vote. Training and inference ran on NVIDIA A100 GPUs using checkpoint_best.pth. Six of 197 validation cases failed on GPU memory and were excluded, leaving 191 evaluable cases.

2.3 Ablation design

To separate data expansion from ensembling, four conditions were evaluated on the same external set: A1 (VerSe-only, fold_all); A2 (VerSe-only, five-fold vote); B (VerSe plus CTSpine1K, fold_all); C (VerSe plus CTSpine1K, five-fold vote). A1-to-B isolates training-data expansion; A1-to-A2 and B-to-C isolate ensembling; the joint comparison indicates which dominates. The VerSe-only baseline used the standard 112-case split, comparable to published VerSe methods.

2.4 Evaluation metrics

Accuracy was quantified with DSC ($2|A \cap B| / (|A| + |B|)$) under two conventions. Our primary metric is the per-vertebra mean DSC — per label, averaged within a case then across cases — appropriate for instance segmentation and stricter than merged foreground. We report an all-labels figure (a label scored whenever present in prediction or reference, excluded only when absent from both) and a presence-matched figure (labels present in both, excluding label-presence artifacts; Section 2.6). The all-labels figure is our headline metric, conservative and directly comparable, following the VerSe challenge [3]. Vertebrae were grouped into cervical (C1-C7), thoracic (T1-T12), and lumbar (L1-L6) regions.

2.5 Ensemble-disagreement reliability analysis

The five-fold ensemble yields five label predictions per voxel. We define per-vertebra fold agreement as the mean fraction of folds assigning the ensemble's majority label within a vertebra's predicted region (1.0 = unanimous). For each vertebra we computed agreement and per-vertebra DSC, then tested whether agreement predicts DSC (Spearman) and whether disagreement discriminates poorly segmented vertebrae (DSC below 0.70) by AUROC, with 95% confidence intervals from 2000 bootstrap resamples over vertebrae. Screening value was assessed as sensitivity against review burden across thresholds.

2.6 Characterization of zero-DSC vertebrae

Vertebrae scoring exactly zero DSC were examined to separate genuine failures from artifacts. For each, we checked whether the reference contained voxels for that label and what each fold predicted, distinguishing label-presence mismatch (predicting a vertebra the reference omits) from true miss (anatomy present but unsegmented). Their distribution across levels was tabulated to test field-of-view clipping.

3. Results

3.1 External segmentation performance and the data-versus-ensemble decomposition

On external CTSpine1K, the expanded five-fold ensemble (C) achieved a mean per-vertebra Dice of 0.9206. The ablation (Table 1) decomposes this: five-fold ensembling of the VerSe-only model raised external Dice from 0.8315 (A1) to 0.8779 (A2), + 4.6 points, and adding a second source raised it to 0.8578 (B), + 2.6 points; combining both was best (C, 0.9206). Neither intervention alone accounts for the full gain, so reports that vary data and inference together cannot attribute improvement to either.

In this task the cheaper intervention was also the more effective one: ensembling (+ 4.6) exceeded a whole second training source (+ 2.6), despite the former reusing folds a standard nnU-Net run already produces. This is specific to the present setting, not a general rule that inference-time aggregation beats data expansion. If anything it is conservative: the training and external sets are disjoint case splits of the same four CTSpine1K cohorts, so part of the data gain reflects distributional alignment rather than generalization. Consistent with a real domain gap, the CTSpine1K authors report a model trained on their annotations dropping from 0.869 in-domain to 0.619 on VerSe [17]; this heterogeneity is what a deployed model must handle.

Table 1
Four-condition ablation on the external CTSpine1K public validation set (191 evaluable cases; 192 for condition A1), computed with a single per-vertebra Dice pipeline across all conditions (per-label Dice averaged within each scan, then across scans; a label is scored whenever present in the prediction or reference, and excluded only when absent from both). Ensembling denotes a five-fold majority vote.

Condition	Training data	Inference	Mean DSC
A1	VerSe only (112)	Single (fold_all)	0.8315
A2	VerSe only (112)	5-fold ensemble	0.8779
B	VerSe + CTSpine1K (387)	Single (fold_all)	0.8578
C (primary)	VerSe + CTSpine1K (387)	5-fold ensemble	0.9206*

* Condition C is the primary configuration. All four conditions were computed by one shared per-vertebra Dice pipeline over the same external split — 6 of 197 cases were excluded for GPU memory, and the VerSe-only single model (A1) completed one more (192) — so they remain directly comparable. The ordering is monotonic: adding a second training source (A1 to B) and adding five-fold ensembling (A1 to A2) each improve external performance, and combining both (C) is best.

3.2 Region-stratified performance and presence-matched reporting

Region-stratified DSC followed the expected anatomical gradient — highest lumbar, lowest cervical (Table 2; presence-matched; representative segmentations in Fig. 1). The all-labels per-case mean of 0.9206 rises to a presence-matched mean of 0.9423 (median 0.9799) once label-presence-mismatch vertebrae are excluded; the difference is entirely the 117 zero-DSC vertebrae (Section 3.4), which are annotation-boundary artifacts, not delineation failures. Performance was consistent across source cohorts (COLONOG 0.9199, MSD-T10 0.9265, COVID-19 0.9197, HNSCC-3DCT-RT 0.8929; all-labels). Lower cervical performance is corroborated by the reliability results below.

Precision and recall closely tracked Dice and were near-balanced (no systematic over- or under-segmentation): presence-matched precision and recall were 0.949 and 0.943 (all-labels 0.921 and 0.916), following the same lumbar-highest to cervical-lowest gradient (cervical 0.888/0.906, thoracic 0.938/0.932, lumbar 0.961/0.955), with 95% confidence intervals from a patient-level bootstrap (2000 resamples over scans).

Table 2
Region-stratified DSC (presence-matched, vertebra-level) for the expanded five-fold ensemble on external CTSpine1K.

Region	Mean DSC	Median DSC	Vertebrae (n)
Cervical (C1-C7)	0.888	0.931	75
Thoracic (T1-T12)	0.933	0.980	1181
Lumbar (L1-L6)	0.957	0.988	906

3.3 Ensemble disagreement predicts segmentation error

Per-vertebra fold agreement was strongly associated with per-vertebra DSC across all 2279 vertebrae (Fig. 2; Spearman 0.862, 95% CI 0.844 to 0.879; $p < 0.001$). Treating DSC below 0.70 as poor (209 of 2279; 9.2%), disagreement discriminated poor from adequate segmentation with AUROC 0.920 (95% CI 0.893 to 0.945); both bootstrap intervals are narrow.

As a screening flag the signal is efficient (Table 3, Fig. 3): referring only vertebrae below 0.8 fold agreement captured 76.1% of poorly segmented vertebrae while flagging just 9.4% overall, at no cost beyond the ensemble already produced during segmentation. Cervical vertebrae showed the lowest agreement and lowest DSC; the folds disagree most precisely where the model is least accurate.

Table 3
Reliability flagging trade-off. A vertebra is flagged when fold agreement falls below the threshold.

Agreement threshold	Sensitivity (poor vertebrae flagged)	Review burden (all vertebrae flagged)
< 0.8	76.1%	9.4%
< 0.7	63.6%	6.7%
< 0.6	50.7%	4.8%

3.4 Characterization of zero-DSC vertebrae

Of 2279 vertebrae, 117 scored exactly zero DSC — not unsegmented visible bone but predominantly label-presence mismatches, where the reference lacked the label so any prediction yields zero overlap. At C1 the reference is often absent because the vertebra is clipped at the superior field-of-view boundary; the CTSpine1K benchmark reports strong C1 segmentation (DSC 0.951) when the label is present [17]. At L6, a transitional variant inconsistently present and annotated, the CTSpine1K baseline itself reports near-zero DSC across partitions [17]. Zero-DSC vertebrae concentrated at C1 (n = 19) and L6 (n = 63), about 70% of the total, mirroring the training-side sliver curation (Section 2.1) — both reflecting inconsistent annotation of partially visible boundary vertebrae. The reliability flag correctly assigns these low confidence, the desired behavior for review.

4. Discussion

This study reframes nnU-Net vertebral segmentation as a question of trust rather than raw accuracy: given a strong but imperfect model, can we say when to believe it, and where its cross-domain performance comes from? Three findings support that aim. First, under genuine cross-domain evaluation, both data expansion and five-fold ensembling improved external performance, with ensembling the larger single-factor gain in this task. This relative magnitude is rarely isolated, since studies typically change data and inference together and report one combined number; computing all four conditions through one pipeline lets a practitioner decide whether to invest in annotated data or aggregate an existing model at inference.

Second, we validate and extend cross-validation-disagreement quality control for spine CT. Gottlich et al. [12] demonstrated training-free fold disagreement at the whole-image level for kidney and tumor segmentation; we confirm it transfers to whole-spine CT and apply it at per-vertebra granularity, with quantified error-predictive validity (AUROC 0.920, bootstrap CIs) and an explicit account of failure modes. Per-structure disagreement has been studied outside the spine [14, 15, 24], so we do not claim the concept as novel; but our units are anatomically indexed instances of fixed identity and count (C1 to L6), so the flag names a vertebra, not variable-count lesions. Following Kirscher et al. [16], who caution that cross-validation folds are not a seed-randomized deep ensemble [13], we test the signal a standard nnU-Net already emits: because folds train on different subsets, their disagreement conflates model

uncertainty with subsampling variability, so we treat it as an empirically validated error tracker, not a calibrated probability. It flags disagreement, not incorrectness, and cannot detect errors on which all folds agree [12].

Third, characterizing apparent failures showed the extreme zero-DSC vertebrae concentrate at the field-of-view boundary (C1, label-presence mismatch) and the transitional L6 level, where mismatch compounds a difficulty the CTSpine1K baseline also encounters [17]. Reporting both all-label and presence-matched metrics makes this transparent and prevents an annotation artifact from being read as a model deficiency. Relative to recent spine CT work, structural uncertainty in myeloma CT [10] and vertebra-aware frameworks [11], our study adds a controlled, whole-spine, externally validated decomposition of performance drivers paired with a bootstrap-quantified, failure-mode-characterized reliability analysis.

Clinically, the reliability flag changes how such a model is used. Rather than trusting or re-checking every level, a radiologist can confine review to the small fraction the flag marks uncertain, inspecting fewer than one in ten vertebrae recovered roughly three-quarters of the poorly segmented ones, cutting manual quality-control time and raising confidence in the vertebrae that downstream fracture grading, morphometry, and opportunistic screening depend on. Natural next steps are to measure that review-time saving prospectively, validate under fully out-of-distribution conditions, calibrate the flag into an actionable probability, and test transfer to other spine tasks and modalities.

Deployment cost is modest, and the reliability signal adds little to it. Running the five folds together in a single inference pass on an NVIDIA A100-PCIE-40GB GPU took approximately 276 s per case, versus 240 s for a single-fold model — only about 15% longer — because per-case runtime is dominated by fixed preprocessing, resampling, and I/O rather than the network forward passes. Peak GPU memory was about 25 GB on typical cases, well within a 40 GB card, though the largest validation volumes exceeded it and account for the six excluded cases. Where even this overhead is undesirable, the flag can be approximated from fewer folds.

4.1 Limitations

The disagreement signal is an uncertainty proxy, not a calibrated probability, and by construction cannot flag the mode in which all five folds are confidently and identically wrong; its correlation with error, though strong, was established empirically on this distribution. Evaluation used the CTSpine1K public validation split, disjoint at the case level from the CTSpine1K training cases but drawn from the same four cohorts, so for the expanded model it is held out but not cross-domain, and the reliability statistics inherit this. Their confidence intervals resample vertebrae rather than patients, unlike the patient-level intervals for precision and recall. Six of 197 cases were excluded for GPU memory. Finally, the presence-matched convention excludes label-presence-mismatch zeros, overwhelmingly annotation artifacts; the all-labels metric is retained as a conservative bound.

4.2 Conclusion

The value of this work lies less in a new segmentation model than in making an existing one trustworthy to deploy. Its central result is that ensemble disagreement, produced for free by a standard nnU-Net workflow, provides a strong per-vertebra reliability signal (AUROC 0.920) that localizes review to the vertebrae most likely to be wrong. Under external validation, five-fold ensembling was also the larger contributor to the cross-domain gain than added data in this task, though we do not generalize that ordering. Reporting both all-label and presence-matched metrics and auditing failure modes separates true error from annotation artifacts. Together these argue for honest, externally validated, decomposed reporting — and for shipping a reliability signal alongside the segmentation — as standards for clinically deployable spine CT tools.

Declarations

Competing interests.

The authors declare that they have no competing interests.

Use of large language models.

During the preparation of this work the authors used a large language model (Claude, Anthropic) to assist with manuscript language editing and with drafting analysis and figure-generation code. All AI-assisted code and text were reviewed, tested, and verified by the authors. No large language model was used to generate, analyze, or interpret the study data, and no AI tool is listed as an author.

Author Contribution

M.F. conceived and designed the study, performed the experiments and analysis, and drafted the manuscript. M.L. and F.V. provided feedback and supervisory guidance, and aided in the revision of the manuscript. Y.T. and M.S. reviewed the paper and provided comments. All authors read and approved the final manuscript.

Data Availability

This study used two publicly available datasets: VerSe (Large Scale Vertebrae Segmentation Challenge) and CTSpine1K. The analysis code supporting this study is available in a public repository.

References

1. Genant HK, Wu CY, van Kuijk C, Nevitt MC (1993) Vertebral fracture assessment using a semiquantitative technique. J Bone Min Res 8(9):1137–1148. 10.1002/jbmr.5650080915

2. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18(2):203–211. 10.1038/s41592-020-01008-z
3. Sekuboyina A, Husseini ME, Bayat A, Löffler M, Liebl H, Li H et al (2021) VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal 73:102166. 10.1016/j.media.2021.102166
4. Löffler MT, Sekuboyina A, Jacob A et al (2020) A vertebral segmentation dataset with fracture grading. Radiol Artif Intell 2(4):e190138. 10.1148/ryai.2020190138
5. Liebl H, Schinz D, Sekuboyina A et al (2021) A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. Sci Data 8:284. 10.1038/s41597-021-01060-0
6. Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 4(3):e210064. 10.1148/ryai.210064
7. Guan H, Liu M (2022) Domain adaptation for medical image analysis: a survey. IEEE Trans Biomed Eng 69(3):1173–1185. 10.1109/TBME.2021.3117407
8. Yang J, Soltan AAS, Clifton DA (2022) Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med 5:69. 10.1038/s41746-022-00614-9
9. Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, Jäger PF (2024) nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024. LNCS vol 15009. Springer, Cham; :488–498. 10.1007/978-3-031-72114-4_47
10. Madzia-Madzou DK, Jak M, de Keizer B, Verlaan JJ, Minnema MC, Gilhuijs KGA (2025) Automated vertebrae identification and segmentation with structural uncertainty analysis in longitudinal CT scans of patients with multiple myeloma. Eur J Radiol 188:112160. 10.1016/j.ejrad.2025.112160
11. Ye ZY, Peng JM, Kamishima T (2026) A vertebra-aware framework for structured analysis of post-fracture lumbar CT. Electronics 15(11):2344. 10.3390/electronics15112344
12. Gottlich HC, Korfiatis P, Gregory AV, Kline TL (2023) AI in the Loop: functionalizing fold performance disagreement to monitor automated medical image segmentation workflows. Front Radiol 3:1223294. 10.3389/fradi.2023.1223294
13. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Adv Neural Inform Process Syst (NeurIPS) ; :6402–6413
14. Roy AG, Conjeti S, Navab N, Wachinger C (2019) Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195:11–22. 10.1016/j.neuroimage.2019.03.042
15. Jungo A, Reyes M (2019) Assessing reliability and challenges of uncertainty estimations for medical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2019. Springer, pp 48–56

16. Kirscher T, Bujotzek M, Kirchhoff Y, Rokuss M, Isensee F, Kahl KC et al (2026) Lost in the folds: when cross-validation is not a deep ensemble for uncertainty estimation. Preprint: arXiv:2605.18329
17. Deng Y, Wang C, Hui Y, Li Q, Li J, Luo S et al (2025) CTSpine1K: a large-scale dataset for spinal vertebrae segmentation in computed tomography. Mach Learn Biomed Imaging 3:824–832. 10.59275/j.melba.2025-gf84
18. Johnson CD, Chen MH, Toledano AY, Heiken JP, Dachman A, Kuo MD et al (2008) Accuracy of CT colonography for detection of large adenomas and cancers. N Engl J Med 359(12):1207–1217. 10.1056/NEJMoa0800996
19. Bejarano T, De Ornelas-Couto M, Mihaylov IB (2018) Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment (HNSCC-3DCT-RT) [Data set]. The Cancer Imaging Archive; 10.7937/K9/TCIA.2018.13upr2xf
20. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA et al (2022) The Medical Segmentation Decathlon. Nat Commun 13(1):4128. 10.1038/s41467-022-30695-9
21. Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z et al (2020) Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 11(1):4080. 10.1038/s41467-020-17971-2
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); :770–778. 10.1109/CVPR.2016.90
23. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015. LNCS vol 9351. Springer, Cham; :234–241. 10.1007/978-3-319-24574-4_28
24. Molchanova N, Raina V, Malinin A, La Rosa F, Depeursinge A, Gales M et al (2025) Structural-based uncertainty in deep learning across anatomical scales: analysis in white matter lesion segmentation. Comput Biol Med 184:109336. 10.1016/j.compbiomed.2024.109336

Figures

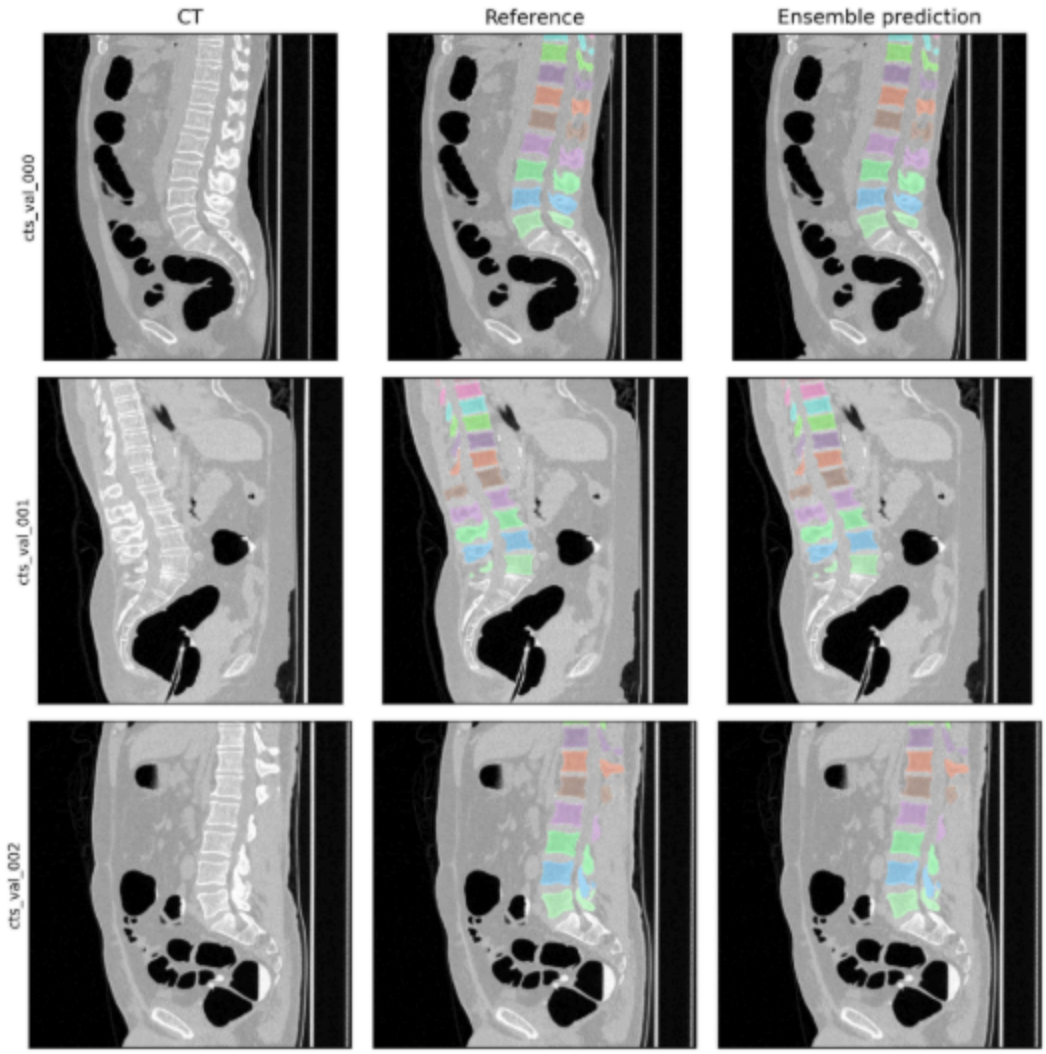


Figure 1

Representative vertebral instance segmentation on the external CTSpine1K validation set. For three cases (rows), a mid-sagittal CT slice (left) is shown with the reference labels (center) and the expanded five-fold ensemble prediction (Condition C, right); colors denote distinct vertebra labels. Predicted and reference labelings correspond closely across the thoracolumbar spine, consistent with the quantitative results in Table 2.

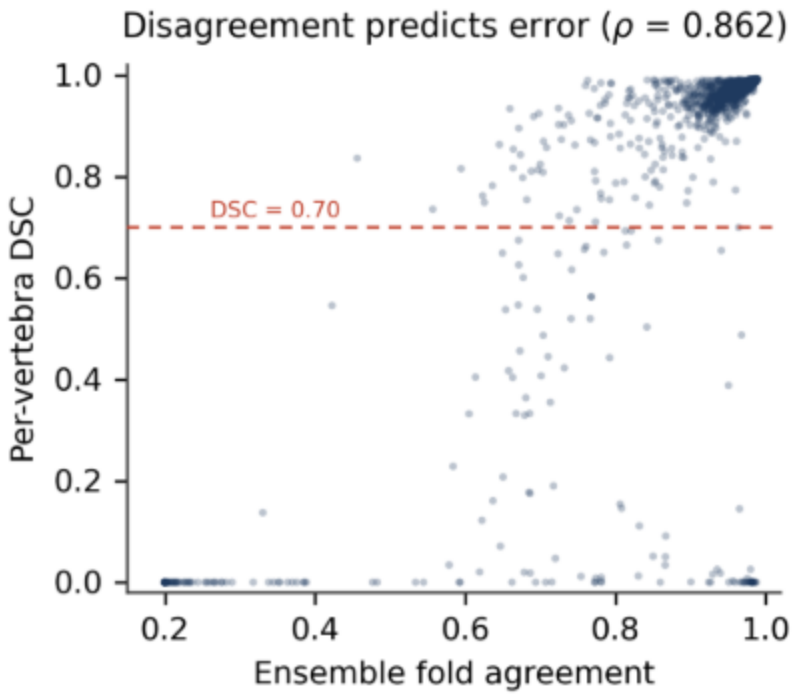


Figure 2

Per-vertebra ensemble fold agreement versus per-vertebra Dice across all 2279 evaluated vertebrae. Lower fold agreement is associated with lower Dice (Spearman rho 0.862, 95% CI 0.844 to 0.879). The dashed line marks the DSC = 0.70 poor-segmentation threshold used for the flagging analysis.

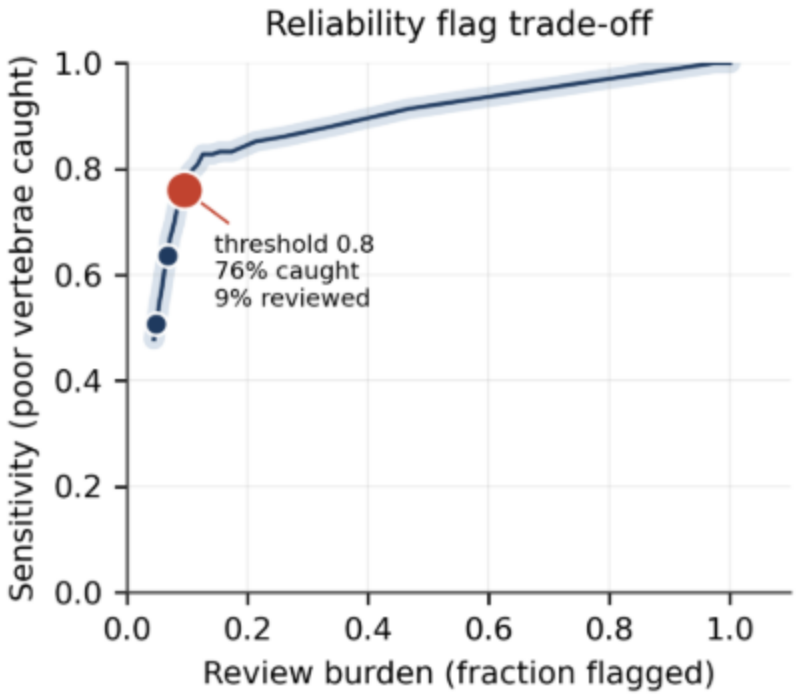


Figure 3

Reliability flagging trade-off. Sensitivity (fraction of poorly segmented vertebrae, DSC < 0.70, that are flagged) against review burden (fraction of all vertebrae flagged) as the fold-agreement threshold varies. Operating points at agreement thresholds of 0.6, 0.7, and 0.8 are marked; the 0.8 threshold captures 76.1% of failures at a 9.4% review burden.