Physics and Imaging in Radiation Oncology 34 (2025) 100740

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Uncertainty-guided pancreatic tumor auto-segmentation with Tversky ensemble

Cenji Yu [a,*], Skylar S. Gay [a], Aashish C. Gupta [a], Rachael M. Martin-Paulpeter [b], Ethan B. Ludmir [c], Yao Zhao [a], Jack Duryea [b], Xinru Chen [a], Carlos E. Cardenas [d], Jinzhong Yang [a,b], Albert C. Koong [c], Tucker J. Netherton [b], Dong Joo Rhee [b], Laurence E. Court [a,b]

[a] The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences (GSBS), 6767 Bertner Avenue, Houston, TX 77030, USA
[b] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA
[c] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA
[d] Department of Radiation Oncology, The University of Alabama at Birmingham, 1700 6th Avenue South, Birmingham, AL 35233, USA

ARTICLE INFO

Keywords:
Auto-segmentation
Pancreatic cancer

ABSTRACT

Background and purpose: Pancreatic gross tumor volume (GTV) delineation is challenging due to their variable morphology and uncertain ground truth. Previous deep learning-based auto-segmentation methods have struggled to handle tasks with uncertain ground truth and have not accommodated stylistic customizations. We aim to develop a human-in-the-loop pancreatic GTV segmentation tool using Tversky ensembles by leveraging uncertainty estimation techniques.

Material and methods: In this study, we utilized a total of 282 patients from the pancreas task of the Medical Segmentation Decathlon. Thirty patients were randomly selected to form an independent test set, while the remaining 252 patients were divided into an 80–20 % training-validation split. We incorporated Tversky loss layer during training to train a five-member segmentation ensemble with varying contouring tendencies. The Tversky ensemble predicted probability maps by estimating pixel-level segmentation uncertainties. Probability thresholding was employed on the resulting probability maps to generate the final contours, from which eleven contours were extracted for quantitative evaluation against ground truths, with variations in the threshold values.

Results: Our Tversky ensemble achieved DSC of 0.47, HD95 of 12.70 mm and MSD of 3.24 mm respectively using the optimal thresholding configuration. We outperformed the Swin-UNETR configuration that achieved the state-of-the-art result in the pancreas task of the medical segmentation decathlon.

Conclusions: Our study demonstrated the effectiveness of employing an ensemble-based uncertainty estimation technique for pancreatic tumor segmentation. The approach provided clinicians with a consensus probability map that could be fine-tuned in line with their preferences, generating contours with greater confidence.

1. Introduction

Radiation therapy is an important pillar in multidisciplinary pancreatic cancer management. Accurate target delineation is crucial for radiation treatment in pancreas to achieve sufficient local control. However, pancreatic tumors are difficult to differentiate from the surrounding parenchyma, even for experienced clinicians. Given the significant level of inherent uncertainty, clinicians rely on clinical intuition to achieve desired level of accuracy in tumor contouring. As a result, tumor contours often have large interobserver variability [1].

In recent years, deep learning-based auto-segmentation has emerged as the preferred method for biomedical image segmentation [2] Although deep learning-based approaches have shown remarkable performance, they often suffer from a tendency towards overconfidence in probability estimation [3]. This can be particularly challenging in segmentation tasks where ground truths are uncertain, as in the case of pancreas tumor segmentation. Models often confidently provide a single erroneous contour upon inference, which requires significant time to

* Corresponding author at: Mayo Clinic Radiation Oncology, 200 First St. SW, Rochester, MN 55905, USA.
E-mail address: yu.cenji@mayo.edu (C. Yu).

https://doi.org/10.1016/j.phro.2025.100740
Received 29 April 2024; Received in revised form 13 February 2025; Accepted 26 February 2025
Available online 8 March 2025

edit and discourages clinical adoption of auto-segmentation. Current deep learning models typically provides a one-size-fit-all solution based on the style of training dataset, allowing little room for customization after inference.

For a segmentation task with uncertain ground-truths like pancreas tumors, providing contours with probability estimates would be ideal. The model, with its prior knowledge from training dataset, would assist experts to create an accurate delineation of the target. To achieve real-world probability estimates using deep learning models, calibration is necessary to ensure the predicted probability map is accurate. Calibration techniques such as Monte-Carlo dropout [4] and test-time augmentation [5] are widely used to generate accurate uncertainty estimates in tandem with the segmentation results, which enables users to obtain interpretable outcomes. However, these calibration techniques require significant time during inference. In addition, these statistical approaches do not accommodate varying segmentation styles that would benefit clinical adoption.

A more efficient approach to address overconfidence in deep learning models is the use of deep ensembles [6]. Fort et al. highlighted that models trained with different configurations can reach their conclusions in distinct ways. By averaging the output probability of high-performance segmentation models within an ensemble, a more robust probability map can be generated that reflects the consensus of expert models. Notably, calibration results are more precise when model configurations diverge [6]. Incorporating a diverse set of model configurations within an ensemble for segmentation provides accurate uncertainty estimation and enhances the model's segmentation performance. The computation required for inferencing would be identical to clinical models, which often leverage multiple models via cross-validation to optimize performance [7].

In this study, we aimed to enhance pancreatic tumor segmentation by conducting deep ensemble calibration on state-of-the-art segmentation models. We hypothesized that the consensus between models with varying segmentation tendencies could efficiently provide accurate pancreatic tumor probability maps and high-quality tumor contours. Furthermore, probability maps would allow clinicians to conduct fast post-hoc adaptations that align with their stylistic preferences. Our approach strived to provide explainable and customizable assistance in pancreatic tumor segmentation and help streamline the clinical workflow.

2. Material and methods

Our study was conducted on the Medical Segmentation Decathlon [8] pancreas tumor task training dataset. A total of 282 contrast-enhanced portal-venous scans collected by the Memorial Sloan Kettering Cancer Center (New York, NY, USA) were used in this study. Scan resolution was 512x512 and slice thickness was 2.5 mm for all patients included in the study. Both pancreatic masses (cyst or tumor) and parenchyma were delineated by an expert abdominal radiologist. To create an independent test set, 30 patients were randomly selected. The remaining 252 patients were divided into a training set (80 %) and a validation set (20 %). To fully leverage the entire training set, we utilized five-fold cross validation. The CT images were clipped from −87 to 199 HU and resampled isotropically at 1.0 mm × 1.0 mm × 1.0 mm. The transformer-based Swin-UNETR architecture was used in this study as shown in Fig. 1. It was selected due to its state-of-the-art performance in the pancreas task of MSD challenge [9]. Given that the Swin-UNETR architecture is 3D-based, we cropped images into 96 × 96 × 96 patches with an overlap of 50 %. Additionally, data augmentation

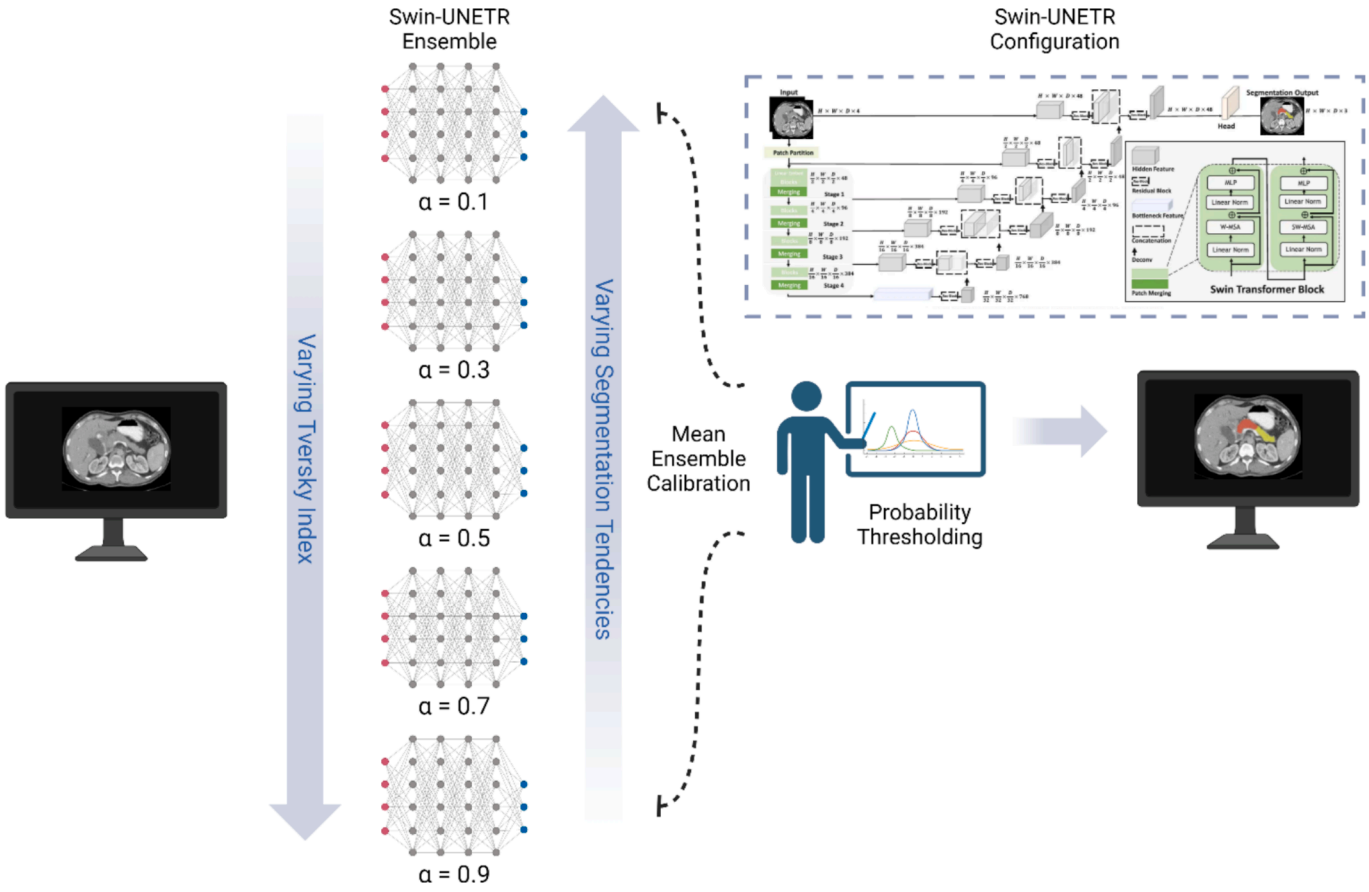

Fig. 1. Swin-UNETR Tversky Ensemble Workflow. The ensemble was created by varying Tversky loss hyperparameter. After inference, the mean of all individual model outputs was calculated to create the probability map. The final contour was created via probability thresholding by experts.

strategies such as random flip, rotation, intensity scaling, and shifting with varying probabilities were employed. The training of the model was conducted on a single A100 GPU for a total of 5000 epochs with a learning rate of $1e^{-4}$ and a batch size of 2. Each member of the ensemble required 151 h to complete training. In order to ensure a fair comparison with the state-of-the-art Swin-UNETR models in the pancreas task, the preprocessing pipeline and hyperparameters were kept identical as reported in Tang et al. [9]. This was done to eliminate any potential confounding factors that could influence the performance comparison.

To integrate various segmentation styles into our ensemble, we utilized the Tversky loss layer during our training process. The baseline ensemble of Swin-UNETR models aims to minimize the Dice similarity coefficient during training, which assigns equal weight to false positives (FP) and false negatives (FN):

$$DSC = \frac{2TP}{2TP + FN + FP}$$

Tversky index, on the other hand, allows us to weigh FP and FN:

$$TI = \frac{TP}{TP + \alpha FN + \beta FP}$$

Here, α and β ($\alpha + \beta = 1$) control the magnitude of the penalties for FN and FP. Through manipulating the Tversky index hyperparameters, we can customize the segmentation tendencies of our models. Models with an α greater than 0.5 have a tendency to under-segment as they penalize false negatives more heavily. Conversely, models with an α less than 0.5 tend to over-segment as they prioritize false positives. However, optimal and well-balanced segmentation is still maximally rewarded regardless of these tendencies.

Utilizing the Tversky loss, we can regulate each model's segmentation tendencies to imitate the contouring styles of multiple experts. We first implemented Tversky loss within the Swin-UNETR architecture with MONAI 1.0 on a voxel-by-voxel basis. Let P be the predicted label from the network and G the ground truth label:

$$TL(P, G; \alpha, \beta) = 1 - TI(P, G; \alpha, \beta)$$

where,

$$TI(P, G; \alpha, \beta) = \frac{|PG|}{|PG| + \alpha|P \backslash G| + \beta|G \backslash P|}$$

During training, α and β controlled the level of penalization for FP and FN respectively. These loss function hyperparameters gave us control in training individual Swin-UNETR model with under-segment or over-segment tendencies. To create a Tversky ensemble, we assigned unique α values to each of the five members. The ensemble was trained with α values of 0.1, 0.3, 0.5, 0.7, and 0.9, respectively.

The model's predictions were generated using sliding windows with a 50 % overlap, and the mean probability of all members in the Tversky ensemble was utilized. The calibrated probability map for tumor prediction was directly extracted from the inference results. A sample workflow of Tversky ensemble auto-segmentation is shown in Fig. 1, where experts utilized the probability map to identify regions that were difficult to visualize on contrast-enhanced CT scans and to correct errors in the auto-segmentation. For quantitative evaluations against the ground truths, we extracted eleven final segmentations by varying the threshold values ranging from 0.05 to 0.9 on the probability map. To accurately reflect segmentation performance after expert thresholding, we selected the contours with the lowest 95th percentile Hausdorff distance to represent the final quantitative results. This approach provided a conservative estimate of performance improvement following expert input. We compared the quantitative performance of the Tversky ensemble, using varying threshold values, against the baseline Swin-UNETR 5FCV ensemble with the Wilcoxon signed-rank test to identify any statistically significant paired differences. This non-parametric test was chosen because the DSC, MSD, and HD95 metrics were non-normally distributed in our experiment. We calculated the sum of ranks for both positive and negative ranks, using the smaller of these two sums as the test statistic. The p-value was then computed to determine whether there was a significant difference between the medians of the Tversky ensemble and the baseline ensemble for the same test patients.

3. Results

Table 1 presents the quantitative results of all the thresholded contours. Our findings indicated that thresholding the probability maps with a value of 0.05 yielded the highest Dice similarity coefficient (DSC) results, while contours thresholded with a probability value of 0.5 exhibited the lowest distance metric.

To mimic the human-in-the-loop adaptation process, we selected the contours with the lowest 95th percentile Hausdorff distance (HD95) among the eleven probability thresholds from each patient for final quantitative evaluation. Our final quantitative results surpassed those of the Swin-UNETR configuration, which achieved state-of-the-art results in the pancreas task of the Medical Segmentation Decathlon challenge. The boxplots of the quantitative results were shown in Fig. 2. Our statistical test has demonstrated that our method not only yielded higher mean and median Dice Similarity Coefficient (DSC) values, but also resulted in lower mean and median distance metrics. This performance improvement was found to be statistically significant for both DSC ($p = 0.006$), HD95 ($p = 0.038$), and average surface distance ($p = 0.012$), with all p-values being less than 0.05 in the Wilcoxon signed-rank test.

When employing an overly cautious segmentation approach (high probability threshold), as depicted in Fig. 3, the generated contours exhibited lower DSC scores, as anticipated. However, we also observed that this conservative contouring strategy led to poorer distance-based results. In contrast, an over-segmentation style yielded more favorable quantitative outcomes in both DSC and distance metrics in pancreatic tumor segmentation.

A sample of the calibrated probability map for the pancreas tumor was shown in Fig. 4. Instead of a fixed contour, the Tversky ensemble produced voxel-by-voxel uncertainty estimation for the tumor. The inference time for the probability map was 18 s.

4. Discussion

In this study, we proposed an approach to address segmentation tasks with uncertain ground truths by utilizing ensemble-based uncertainty estimation techniques. Deep ensembles have demonstrated remarkable performance in uncertainty estimation tasks, and greater variability within the ensemble has been observed to improve the calibration of the pixelwise probability map [6]. To introduce human-like variability and incorporate multiple segmentation styles into the

Table 1
Quantitative Analysis of contours created with varying thresholds of the probability map. Best results were created from selecting the contours with the lowest HD95 for each individual case. The Swin-UNETR results were from a 5fcv Swin-UNETR ensemble trained with DSC loss.

	DSC		HD95 (mm)		MSD (mm)	
	Mean	SD	Mean	SD	Mean	SD
0.05	0.47	0.33	14.4	14.2	4.9	9.5
0.1	0.46	0.33	14.8	14.5	4.9	9.6
0.2	0.45	0.33	14.9	14.6	5.0	9.6
0.3	0.44	0.34	15.4	15.9	5.7	10.6
0.4	0.44	0.34	15.6	16.1	5.6	10.5
0.5	0.43	0.34	14.0	11.0	4.1	6.1
0.6	0.42	0.34	14.1	11.0	4.1	6.1
0.7	0.41	0.34	14.2	11.0	4.1	6.1
0.8	0.40	0.33	14.3	11.1	4.1	6.1
0.9	0.39	0.33	14.5	11.1	4.1	6.1
Best	0.47	0.33	12.7	10.0	3.2	4.6
Swin-UNETR	0.43	0.34	13.4	10.1	3.8	4.1

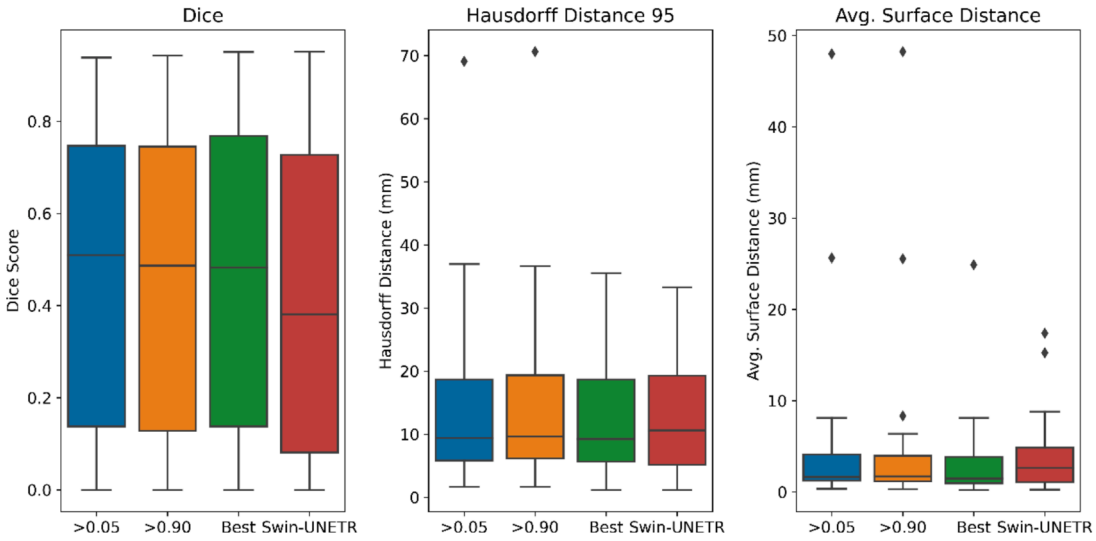

Fig. 2. Quantitative results of automatically generated contours compared to ground truths. Contours were generated by thresholding the probability map with a variety of values (0.05 and 0.9 as shown) and the contour with the lowest HD95 were chosen to serve as the best contour to compare against the Swin-UNETR ensemble.

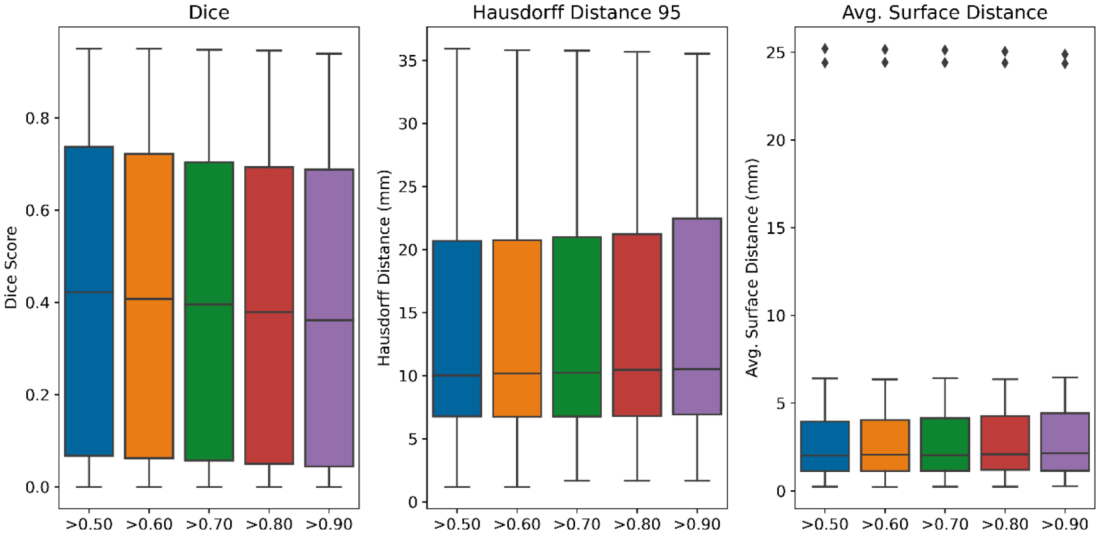

Fig. 3. The trend of segmentation quality as thresholding value increases was shown in boxplots. The DSC scores between generated contours and ground truths decreased and the distance metric increased.

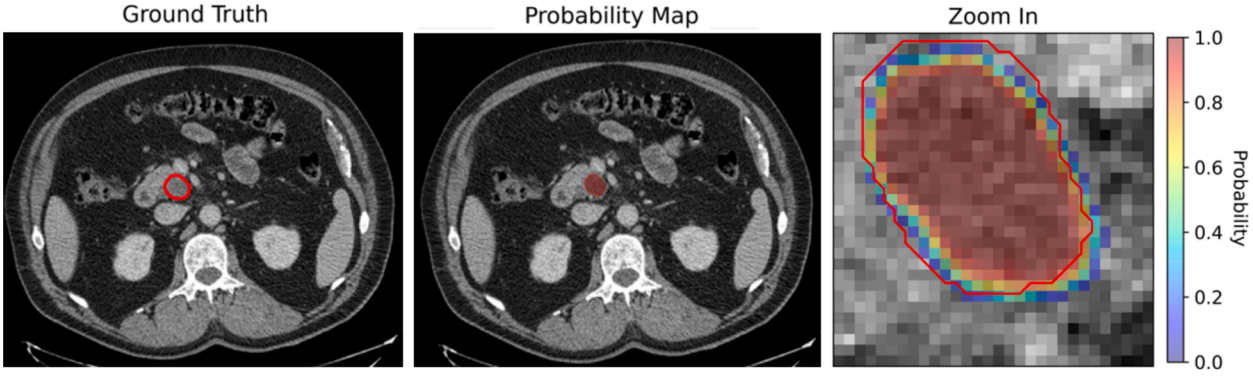

Fig. 4. A probability map generated by Tversky ensemble. Final segmentations were derived from thresholding the probability map.

consensus probability estimation, we employed the Tversky loss function to fine tune the contouring style of each individual model [10]. In addition to using different data folds, we tuned the Tversky hyperparameters to generate models with varying segmentation tendencies. This enabled the creation of multiple segmentations from a well-calibrated probability map that can be adjusted to the physician's preferences as shown in Fig. 4. Our approach yielded superior quantitative results compared to the Swin-UNETR [11] ensemble, which was trained and tested on the same dataset with identical cross-validation data folds. Both the Tversky ensemble and the Swin-UNETR ensemble were trained using the preprocessing and hyperparameters reported by the ensemble that achieved state-of-the-art performance in the pancreas task of the Medical Segmentation Decathlon [8].

The Dice similarity coefficient (DSC) results showed consistent improvement with lower probability thresholds as shown in Fig. 3. Upon qualitative observation, we found that the model consistently under-segmented the tumor compared to the ground truth. While the generated segmentations captured the hypodense regions in the CT images, they failed to extrapolate to the surrounding diseased areas that were less prominent to the human eye. By lowering the probability threshold, the generated contours became more aggressive in delineating the uncertain regions at the tumor border. This resulted in a greater overlap with the ground truths labeled by experts, as depicted in Fig. S1, leading to improved quantitative performance. Selecting the contours based on the lowest HD95 distance further improved the distance metrics without compromising the DSC. By optimizing the thresholding strategy on a patient-by-patient basis, we retained aggressive segmentations that incorporated uncertain regions while eliminating erroneous regions with low confidence. In the clinical workflow, physicians could threshold the probability map in real-time to accommodate their preferences. The post-hoc editing capability introduced expert input and allowed fast adaptation prior to contour finalization. The pixel-level uncertainty estimation also offered clinicians more confidence when identifying the pancreatic tumor target, especially at the tumor border. Our Tversky ensemble enhanced the auto-segmentation workflow for clinicians, allowing fast post-hoc adaptation and provided contouring assistance at tumor border.

While over-segmentation was preferred in pancreatic tumor segmentation due to the inherent uncertainty at tumor borders, incorporating low probability regions was not without its drawbacks. When lenient thresholding was applied, the ensemble could falsely identify tumors from benign anatomy, as illustrated in Fig. S2. This occurrence was common in auto-segmentation since pancreatic tumors often displayed low contrast compared to the surrounding tissue. False positives were frequently observed due to the presence of hypodense regions throughout the CT scans. In our post-processing step, we retained the largest connected component of the predicted contours, which could result in falsely labeled low probability regions becoming the larger connected component and leading to poor quantitative results. This perturbation to the distance metrics occurred when increasing the threshold from 0.4 to 0.5. However, the calibrated probability map offered an opportunity to detect some mis-contoured cases based on uncertainty estimates. Clinicians could visually identify regions with low confidence. If the initial probability map was found to be erroneous, they had the ability to eliminate falsely identified tumor regions by increasing the probability threshold, as demonstrated in Fig. S2. This feature allowed the model to maintain an aggressive approach in most cases to ensure optimal results, while producing accurate contours after human intervention when the Tversky ensemble was uncertain.

Based on our observation that over-segmentation tendencies yielded contours that were closer to the ground truth, we selected a Tversky α value of 1.0 to construct an ensemble that maximally rewarded over-segmentation. We utilized identical data split, preprocessing techniques, and hyperparameters outlined by the state-of-the-art Swin-UNETR ensemble. The resulting over-segmenting ensemble only achieved an average DSC of 0.40. This performance deterioration underscored the significance of the diversity introduced by the varying Tversky hyperparameters. It substantiated the crucial role of a well-calibrated probability map in achieving accurate segmentation. Directly tuning the model towards the desired behavior did not yield improvements in segmentation quality. By incorporating diverse segmentation tendencies within the Tversky ensemble, we successfully generated a probability map that was better calibrated along with more precise segmentation.

Our proposed approach still required human intervention for the final contouring. The state-of-the-art approach achieved an average DSC of 0.43 in our test set, indicating that expert input remained necessary for achieving optimal plan quality in segmentation workflows. Despite outperforming the state-of-the-art model ensemble using identical preprocessing and hyperparameters, our tool was not yet capable of fully automating the segmentation of pancreatic tumors. In addition, while our Tversky ensemble improved upon baseline ensemble in under-segmenting, the model was still not sufficiently aggressive in uncertain regions. Additionally, our post-processing pipeline, which retained the largest component, might introduce unintended variabilities when using low probability thresholds for aggressive contouring at the tumor border. Therefore, caution is advised when conducting thresholding to avoid compromising the accuracy and quality of the final segmentation results.

One limitation of our study was the tendency of supervised deep learning segmentation models to under-segment pancreatic tumors. The segmentation performance was particularly affected at tumor borders where experts relied on clinical notes in conjunction with imaging features. Recently, text-guided segmentation has been introduced in medical imaging. Text prompts have proven effective in guiding transformer-based segmentation networks in areas where imaging features alone are insufficient for vision-only deep learning models [12]. For future work, we plan to integrate natural language guidance into our vision transformer-based segmentation network. By utilizing information from clinical notes, we aim to minimize under-segmentation at tumor borders and achieve expert-level performance in pancreatic tumor segmentation. Generating segmentations from clinical notes can also help reduce inter-observer variability in clinical interpretation and facilitates the creation of standardized datasets.

The current study was also limited by the lack of qualitative assessment by experts. While our experts assisted in finalizing the workflow with the Tversky ensemble, the resources and cost required for a large-scale multi-institution qualitative evaluation prohibited us from including it in this study. Future work will focus on deploying the model in clinical settings and conducting qualitative assessments of the contours with multiple experts from various institutions. While the primary challenge in pancreatic tumor segmentation lies in uncertain ground truths, the qualitative evaluation will create ground-truth contours from different institutions. This diverse dataset will be utilized to enhance the performance of pancreas tumor segmentation models and improve their generalizability. By combining text-guided segmentation with an extensive qualitative review of a diverse dataset, we hope to achieve fully automated pancreatic tumor segmentation.

In this study, we employed an ensemble-based uncertainty estimation technique to facilitate the segmentation of pancreatic tumors. Given the inherent ambiguity of ground truth delineation, we adapted the Tversky loss function to account for a variety of contouring styles and generate a consensus probability map that can be fine-tuned by clinicians in line with their preferences, following model inference. By utilizing the same network architecture, data preprocessing pipeline, hyperparameters, and ensembling strategy as the state-of-the-art model, our approach outperformed its Swin-UNETR counterpart in the pancreatic tumor segmentation task of the Medical Segmentation Decathlon. Furthermore, our method provides pixel-wise uncertainty estimation, which enables clinicians to generate contours with greater confidence. We are optimistic that our Tversky ensembles can serve as an accurate and dependable solution for pancreatic tumor segmentation.

CRediT authorship contribution statement

Cenji Yu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Skylar S. Gay:** Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Aashish C. Gupta:** Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Rachael M. Martin-Paulpeter:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing. **Ethan B. Ludmir:** Conceptualization, Investigation, Resources, Validation, Data curation, Writing – review & editing. **Yao Zhao:** Conceptualization, Investigation, Resources, Validation, Data curation, Writing – review & editing. **Jack Duryea:** Methodology, Software, Investigation, Writing – review & editing. **Xinru Chen:** Methodology, Software, Investigation, Writing – review & editing. **Carlos E. Cardenas:** Conceptualization, Methodology, Writing – review & editing. **Jinzhong Yang:** Conceptualization, Resources, Writing – review & editing. **Albert C. Koong:** Conceptualization, Resources, Validation, Writing – review & editing. **Tucker J. Netherton:** Supervision, Methodology, Validation, Writing – review & editing. **Dong Joo Rhee:** Supervision, Methodology, Validation, Writing – review & editing. **Laurence E. Court:** Conceptualization, Methodology, Project administration, Funding acquisition, Supervision, Resources, Writing – review & editing.

Declaration of competing interest

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

Acknowledgements

This work was supported by the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program (STRIDE). The authors acknowledge the support of the High Performance Computing for research facility at the University of Texas MD Anderson Cancer Center for providing computational resources that have contributed to the research results reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2025.100740.

References

[1] Wong J, Baine M, Wisnoskie S, Bennion N, Zheng D, Yu L, et al. Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-95152-x.
[2] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol 2019;92(1100):20190001. https://doi.org/10.1259/bjr.20190001.
[3] Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging 2020;39:3868–78. https://doi.org/10.1109/TMI.2020.3006437.
[4] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning 2015.
[5] Wang G, Li W, Aertsen M, Leuven KU, Deprest J, Ourselin S, et al. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation. n.d.
[6] Fort S, Hu H, Lakshminarayanan B. Deep ensembles: a loss landscape perspective 2019:1–15.
[7] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.
[8] Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. Nat Commun 2022;13. https://doi.org/10.1038/s41467-022-30695-9.
[9] Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, et al. Self-supervised pre-training of swin transformers for 3D medical image analysis. CVPR 2022:20730–40. https://doi.org/10.1109/cvpr52688.2022.02007.
[10] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks 2017.
[11] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D, et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv preprint arXiv:2201.01266; 2022.
[12] Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, et al. LViT: language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging 2024;43:96–107. https://doi.org/10.1109/TMI.2023.3291719.