Segmentation of Pancreatic Ductal Adenocarcinoma using nnU-Net ResEnc with Tversky Loss

Group 19, PANORAMA

Giedrius Mirklys
s1101773
giedrius.mirklys@ru.nl

Ignas Golubovskis
s1036322
ignas.golubovskis@ru.nl

Björn Westerlund
s1158123
bjoern.westerlund@ru.nl

Luka Godnič
s1156065
luka.godnic@ru.nl

Abstract—**Pancreatic ductal adenocarcinoma (PDAC) is one of the deadliest types of pancreatic cancer, often eluding early detection due to its subtle presentation on contrast-enhanced CT (CECT) scans. The PANORAMA challenge provides a large-scale, multi-reader benchmark for PDAC detection, offering a unique opportunity to evaluate and enhance machine-learning–based diagnostic tools. In this project, we aimed to improve upon the winning submission of the PANORAMA competition, which leverages the nnU-Net framework for segmentation and classification. Our focus was on reducing false positives to minimize radiologist workload, primarily by incorporating the Tversky loss function, which allows fine-tuning the trade-off between false positives and false negatives. [A hint on the results]**

We analyzed the effect of this modification on detection accuracy and the area under the ROC curve (AUC), and further evaluated segmentation performance using Dice score metrics. This report details our methodological improvements, experimental results, and future directions for robust, generalizable PDAC detection models [might be excluded later].

I. Introduction

The PANORAMA grand challenge [4] is the first large-scale reader study designed to establish baseline radiologist performance in detecting Pancreatic Ductal Adenocarcinoma (PDAC) on contrast-enhanced computed tomography (CECT) scans. PDAC is the most common type of malignant tumor affecting the pancreas and is among the deadliest of all solid cancers. In the United States alone, approximately 67,440 new cases and 51,980 deaths from PDAC are projected for 2025 [2].

PDAC often leaves only faint radiologic footprints—e.g. minimal ductal calibre changes or subtle parenchymal texture shifts—on contrast-enhanced CT (CECT) months before a tumour becomes obvious. These patterns are easily missed by experts reviewing thousands of abdominal scans for diverse indications. Modern deep-learning systems, however, thrive on precisely this kind of weak-signal problem: by optimizing millions of parameters across 3-D feature hierarchies, they can amplify and aggregate minute cues that fall below human perceptual thresholds [6], [7]. See a sample CECT scan from the Grand Challenge dataset in Figure 1, marking the lesion location.

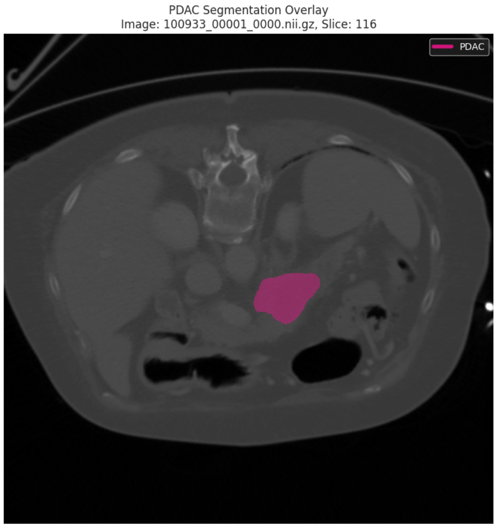
PDAC Segmentation Overlay
Image: 100933_00001_0000.nii.gz, Slice: 116
PDAC

Fig. 1. A CECT scan slice illustrating Pancreatic Ductal Adenocarcinoma (PDAC).

The challenge involved over 68 international radiologists and a hidden test cohort of more than 400 cases. The reader study consisted of two components: (1) radiologists provided a binary diagnosis along with a PDAC likelihood score, and (2) for cases identified as PDAC, they marked the lesion location using point annotations. This setup enabled a rigorous comparison between human and algorithmic performance. The PANORAMA challenge thus provides the first large-scale test bed for PDAC detection systems, with several distinguishing characteristics:

- **Data volume.** Annotated scans of >2238 cases for open development.
- **Human baseline.** 68 international radiologists supply a reader-study benchmark, enabling rigorous algorithm-vs-expert comparison.
- **Task design.** Algorithms are evaluated on (i) binary PDAC presence scores and (ii) voxel-level lesion localisation, mirroring real-world deployment needs.

State-of-the-art deep learning approaches, such as the PANDA model described by Cao et al. (2023) which utilizes components like CNNs and transformers, have reported NCCT AUCs > 0.95 for pancreatic lesion detection on single-centre cohorts [8]. Yet external validation is rare, and most models are tuned on datasets an order of magnitude smaller than

PANORAMA. Critically, these smaller datasets typically have fewer different data sources, leading to lower overall variance between samples, which will likely boost prediction scores but is unlikely to generalize well.

In this project, we aimed to improve upon the winning submission of the PANORAMA competition [reference to Hiu et al] by focusing on reducing false positives (FP) in the detection of PDAC. We explored a single approach to achieve this: modifying the loss function used during training to incorporate the Tversky loss [9], which allows for a more nuanced control over false positive and false negative rates. For example, [9] showed that incorporating the Tversky loss in their 3D-UNet model for detecting multiple sclerosis improved the F2 score from 51.77 to 57.32, compared to the Dice loss alone. Additionally, reducing FP rates is particularly relevant in medical imaging, where the cost of false positives can lead to unnecessary additional tests and increased workload for radiologists. We also compared to the baseline model, which used a standard cross-entropy loss function [citation], and evaluated the impact of this modification on the area under the ROC curve (AUC) and the Dice score, showing the segmentation performance of the model.

II. METHODS

This section outlines the methods used to improve the PDAC segmentation model, focusing on the modifications made to the baseline model and the rationale behind these changes. We also elaborate on the dataset used, Tversky loss, and experimentation.

A. Dataset

The dataset used in this project is the PANORAMA dataset, which consists of 2238[check the number] contrast-enhanced computed tomography (CECT) scans annotated by radiologists. The dataset was not split a priori into training and test sets, as the competition organizers had a separate hidden test set for evaluation on the Grand Challenge platform. However, the nnUNet framework trains the models in folds, where each fold is split into training and validation sets anew. Table **??** and Figure **??** show the distribution of PDAC presence in the dataset, indicating that the dataset is imbalanced, with a significant number of cases without PDAC.

B. Baseline Model and nnU-net Framework

We chose to use the winning submission of the PANORAMA competition as a starting point for our own experiment. We considered using the competition baseline but found them to be very similar, with the competition winner's code repository being easier to follow. This approach uses nnU-net [1], an out of the box neural network solution that builds on the U-net architecture that is specifically tuned for biomedical use cases. It uses a provided plan, often described in a `.json` file, to apply hyper-parameters and allows for automatic hyper-parameter optimization for optimal performance on a given dataset. These include input dimensions (2D vs. 3D), image pre-processing, a neural network architecture, and post-processing.

The full inference pipeline consists of 4 steps: down-sampling, low-res prediction, masking, and high-res prediction. First, the CECT images are loaded, and their voxels are down-sampled according to a set spacing, primarily to manage computational resources and standardize input, saving computational cost and without losing predictive ability. Then the nnU-net makes a mask prediction, aiming to identify the rough area of the pancreas. The mask is represented as a bounding box which is applied to the original image. Finally, a second prediction is made: at the full resolution within the bounding box, where the pancreas should be located. This prediction produces two outputs, the full scan of the patient with lesion candidate likelihoods applied to each voxel as well as a single number representing the probability that a patient has a pancreatic lesion. Here, the Cross-Entropy is explicitly set for the loss.

The network architecture itself of the winning submission has no differences in the first prediction, for identifying the pancreas. However, unlike the baseline, a residual encoder U-net is used in the lesion segmentation. This means that the convolution blocks on the contraction side of the U-net have skip connections like those used in the original ResNet architecture [11]. The purpose of these skip connections is to better support deeper networks by avoiding the vanishing gradient problem. According to a recent paper by the maintainers of nnU-net, this approach achieves state of the art performance on 3D medical imaging tasks, beating newer architectures like transformers or mamba [10]. As a final note, the nnU-net has an implementation for both a 2D and 3D approach, but the 3D one is preferred for this task, as it was also chosen by the authors of the baseline model, and it allows the model to learn from the spatial context of the CECT scans, which is crucial for accurately identifying PDAC lesion.

III. IMPROVING THE BASELINE WITH TVERSKY LOSS

An effective strategy for improving the baseline model is to reduce the false positive (FP) rate, thereby decreasing the number of cases requiring unnecessary review by radiologists and physicians. In this context, shifting the area under the ROC curve (AUC) to the left is desirable, as it corresponds to a lower FP rate at clinically relevant sensitivity levels. This approach directly addresses the practical need to minimize radiologist workload while maintaining high detection performance.

One approach we investigated involved training the network using the Tversky loss, a loss function widely adopted in medical image analysis for its ability to handle class imbalance effectively. The Tversky loss quantifies the dissimilarity between the predicted segmentation and the ground truth, and can be viewed as a generalization of the Jaccard index [3]. It introduces two tunable parameters, α and β, which allow differential penalization of false positives (FP) and false negatives (FN). Specifically, setting $\alpha > \beta$ increases the penalty on false positives, while $\beta > \alpha$ emphasizes reducing false negatives. This flexibility allows Tversky loss to fine-tune the trade-off between precision and recall, making it particularly useful in scenarios where the cost of false positives and false negatives is not equal. In our case, where minimizing false positives and maximizing precision is crucial, Tversky loss provides an effective way to bias the model accordingly.

The Tversky index (TI) is defined as:

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

where TP, FN, and FP denote the number of true positives, false negatives, and false positives, respectively. The Tversky loss is then computed as $1 - \mathrm{TI}$. By adjusting the values of α and β in the equation above, we can control the trade-off between penalizing false positives and false negatives during training.

To integrate the Tversky loss into the nnU-Net framework, we implemented a custom trainer class, `nnUNetTrainerTverskyLoss`, which inherits from the base nnUNetTrainer. This custom trainer was modified to utilize the Tversky loss function, with parameters $\alpha = .3$ and $\beta = .7$, for the segmentation task during model training, replacing the default Cross-Entropy loss for this component. Owing to significant computational requirements, the model incorporating Tversky loss was trained and evaluated on a single data fold from the nnU-Net's 5-fold cross-validation scheme. The baseline Cross-Entropy model was trained on all 5 folds for comparison. Performance metrics reported in Section IV for the Tversky model are based on this single fold. For comparative analysis, Receiver Operating Characteristic (ROC) curves were generated for both the baseline and Tversky loss models based on their patient-level PDAC probability scores. Furthermore, qualitative comparisons of segmentation predictions were performed on identical samples from the test set of the evaluated fold.

IV. Results

We trained our model on the Radboud cluster for 298 epochs, and we chose to use a single fold as this was highly computationally intensive. In figure 2 we can see the loss and dice progression over the course of the training.

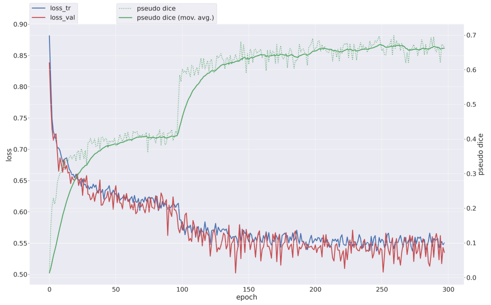

Fig. 2. Progression of loss and psuedo dice values by epoch

Based on the graph, the number of epochs seems reasonable as both loss and dice seem to plateau after around 250 epochs. Looking at only the loss, the variance between epochs appears quite large, especially on the validation data. But since the training loss is not quite as volatile and the moving average of the pseudo dice score is consistently increasing, we did not find this worrying enough to warrant a re-run with a lower learning rate.

An interesting observation is that the Dice score increases sharply around the 100th epoch, while the corresponding decrease in loss is comparatively modest. The underlying cause of this phenomenon is not immediately clear and may warrant further investigation.

V. Conclusions

When building any machine learning models for medical applications, it is important to consider not only the performance metrics such as accuracy or F_1 score but also the practical implications of false positives and false negatives. In this project, we focused on reducing false positives in the detection of pancreatic ductal adenocarcinoma (PDAC) by modifying the loss function used during training to incorporate the Tversky loss. This approach allowed us to fine-tune the trade-off between false positives and false negatives, which is particularly relevant in medical imaging where the cost of false positives can lead to unnecessary additional tests and increased workload for radiologists.

We implemented a custom trainer class in the nnU-Net framework to utilize the Tversky loss function, and trained the model on a single fold of the PANORAMA dataset. The results showed that the Tversky loss model achieved a [performance] area under the ROC curve (AUC) compared to the baseline model using cross-entropy loss, indicating [...] performance in reducing false positives. Additionally, qualitative comparisons of segmentation predictions demonstrated that the Tversky loss model produced [....] accurate segmentations of PDAC lesions.

Our findings are in line with previous studies that have shown the effectiveness of Tversky loss in improving segmentation performance in medical imaging tasks. However, yet alone, the Tversky loss may not be optimimal for decreasing false positives, and other its variations tend to be more effective, such as the Focal Tversky loss [3]. Future work could explore the use of Focal Tversky loss or other variations to further improve the model's performance in reducing false positives. Additionally, we could also consider incorporating attention mechanisms or other advanced techniques to enhance the model's ability to focus on relevant features in the CECT scans.

References

[1] Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

[2] National Cancer Institute, "SEER Cancer Stat Facts: Pancreatic Cancer," https://seer.cancer.gov/statfacts/html/pancreas.html, accessed May 15, 2025.

[3] Saba, M., "Tversky Loss," *Medium*, https://medium.com/@saba99/tversky-loss-902f5f8cc35f, accessed May 15, 2025.

[4] PANORAMA Grand Challenge, "PANORAMA: AI Grand Challenge for Pancreatic Cancer Detection," https://panorama.grand-challenge.org/, accessed May 15, 2025.

[5] The Radiology Assistant, “Pancreas - Carcinoma,” https://radiologyassistant.nl/abdomen/pancreas/pancreas-carcinoma-1, accessed May 15, 2025.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[7] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, A. Chesus, D. C. Corrado, B. Darzi, *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, no. 7788, pp. 89–94, 2020, doi: 10.1038/s41586-019-1799-6.

[8] K. Cao, Y. Xia, J. Yao, H. Xu, L. Lambert, T. Zhang, W. Tang, G. Jin, J. Hui, F. Xu, I. Nogues, X. Li, W. Guo, Y. Wang, W. Fang, M. Qiu, Y. Hou, T. Kovarnik, M. Vocka, Y. Lu, Y. Chen, X. Chen, Z. Liu, Z. Chen, C. Xie, R. Zhang, H. Lu, G. Hager, A. L. Yuille, L. Lu, C. Shao, Y. Shi, T. Liang, L. Zhang, and J. Lu, “Large-scale pancreatic cancer detection via non-contrast CT and deep learning,” *Nature Medicine*, vol. 29, no. 12, pp. 3033–3043, 2023, doi: 10.1038/s41591-023-02640-w.

[9] Salehi, S.S.M., Erdogmus, D., Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: Wang, Q., Shi, Y., Suk, HI., Suzuki, K. (eds) Machine Learning in Medical Imaging. MLMI 2017. Lecture Notes in Computer Science(), vol 10541. Springer, Cham. https://doi.org/10.1007/978-3-319-67389-9_44

[10] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F. (2024). nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv preprint arXiv:2404.09556. https://arxiv.org/abs/2404.09556

[11] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385. https://arxiv.org/abs/1512.03385