Segmentation of Pancreatic Ductal Adenocarcinoma using ResUNet with Tversky Loss

Group 19, PANORAMA

Giedrius Mirklys
s1101773
giedrius.mirklys@ru.nl

Ignas Golubovskis
YOUR$_S TUDENT_N UMBER$
ignas.golubovskis@ru.nl

Björn Westerlund
YOUR$_S TUDENT_N UMBER$
bjoern.westerlund@ru.nl

Luka Godnič
YOUR$_S TUDENT_N UMBER$
luka.godnic@ru.nl

***Abstract*—Pancreatic ductal adenocarcinoma (PDAC) is among 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 [**?**] 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 [**?**].

Pancreatic ductal adenocarcinoma (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 [**?**], [**?**]. See a sample CECT scan from the Grand Challenge dataset in Figure **??**, marking the lesion location.

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:

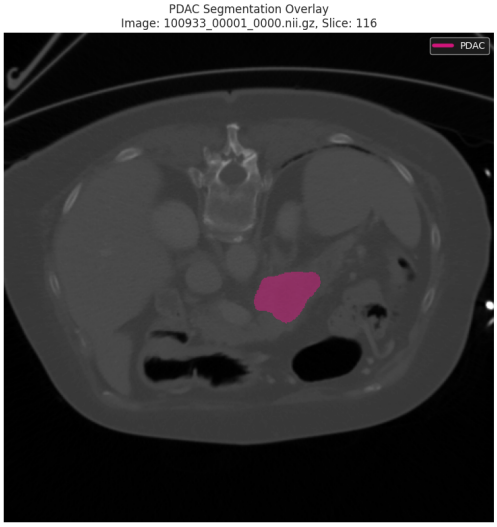

Fig. 1. A CECT scan slice illustrating Pancreatic Ductal Adenocarcinoma (PDAC).

- **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 [**?**]. 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 [**?**], which allows for a more nuanced control over false positive and false negative rates. For example, [**?**] 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.

A. Model starting point

We chose to use the winning submission of the PANORAMA competition as starting point for out own experiments. We considered using the competition baseline but found them to be very similar, with the competition winner's repo being easier to follow. This approach uses nnU-net, [**?**] an out of the box neural net solution that builds on the U-net architecture that is specifically tuned for biomedical use cases and contains. Additionally, nnU-net can automatically adapt preprocessing, neural net architecture, and post processing according to the data it is classifying.

The full inference pipeline consists of 4 steps: down-sampling, low-res prediction, masking and high-res prediction. First images are loaded and the voxels are down-sampled according to a set spacing, main to save. Then the nnU-net makes a mask prediction, aiming to identify the rough area of the pancreas. Size-standardized down-sampling save computational cost without losing predictive ability. All hyper-parameters for the prediction is left to the nnU-net defaults. The mask is represented as a bounding box which is applied to the original image. Finally, a second prediction is made, this time at full resolution but only using the area within the bounding box. 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, with the rest of the

II. IMPROVING THE BASELINE

When considering how to improve the baseline, we explored various potential approaches and evaluated their feasibility. We quickly determined that the most effective objective was to shift the AUC curve to the left. This shift reduces the false positive (FP) rate, thereby decreasing the workload for radiologists and physicians, who must review all cases flagged as positive.

A. Tversky loss

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 [**?**]. 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.

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

There are a few more things we need to describe in the report:

- checking the data: balance and differences in test and training data - besides the loss, dice score curves of both baseline and Tversky, maybe some outputs of the filters
- some FP/FN analysis.

III. RESULTS

TODO

IV. CONCLUSIONS

TODO

This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

***Index Terms*—component, formatting, style, styling, insert**

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$_4$4