Multi-task Pancreas Cancer Segmentation and Classification with nnU-Net V2

David Guo
Divison of Engineering Science
University of Toronto
Toronto, Canada
davidmy.guo@mail.utoronto.ca

***Abstract*—Pancreatic cancer is a highly lethal malignancy, with survival rates heavily dependent on timely and accurate diagnosis. Subtype classification and tumor segmentation from medical images are critical for guiding treatment strategies. In this work, we present a multi-task model based on the nnU-Net V2 framework, designed to perform simultaneous segmentation of pancreatic tumors and classification of their subtypes. The model employs a shared residual encoder and dual decoder architecture, with a custom classifier head for subtype prediction.**

We evaluated the model on a de-identified pancreatic computed tomography (CT) dataset using a 5-fold cross-validation approach. The classifier achieved an accuracy of 91.67%, with average sensitivity and specificity of 90.4% and 91.4%, respectively, demonstrating performance comparable to human experts. The segmentation head attained a Dice score of 45.36 and an Intersection over Union (IoU) of 35.20%, indicating challenges in segmenting complex tumor regions.

***Index Terms*—Pancreatic cancer, Medical image segmentation, nnU-Net, encoder-decoder, Multi-task learning**

I. Introduction

Pancreatic cancer is one of the deadliest malignancies, with limited survival rates due to late diagnosis and complex treatment pathways. Moreover, with an increasing incidence rate, it is predicted to become the second leading cause of cancer-related death in several regions, including the United States and Europe [1]. Accurate identification and classification of pancreatic lesions and their subtypes in CT scans are critical for improving diagnostic accuracy and treatment planning. Prognosis and treatment for differing subtypes varies greatly, from a median survival of 13.3 months for Squamous PDAC types up to 30 months for ADEX [2]. Moreover, the majority of PDAC patients are not eligible for surgical tumor sampling, further emphasizing the need for non-invasive subtype diagnosis. Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated significant success in medical imaging tasks, including segmentation and classification [3].

The nnU-Net framework [4] has emerged as a state-of-the-art platform for medical image segmentation, offering automated pipelines that adapt to diverse datasets and tasks. In this work, we extend the capabilities of nnU-Net V2 to tackle a multi-task problem: simultaneous segmentation of pancreatic tumors and classification of their subtypes. Leveraging a shared encoder and dual decoders, our model aims to streamline the workflow by combining these tasks within a single architecture.

We present a thorough investigation of our model, including architectural modifications to incorporate a classification head, loss functions, and training strategies. Our approach is evaluated using a de-identified pancreatic CT dataset, with performance metrics assessed through 5-fold cross-validation and ensemble predictions. Additionally, we perform ablation studies to analyze the contributions of various architectural components to classification performance.

II. Dataset

The data set is non-public and consisted of de-identified pancreatic CT scans, labeled with background, pancreas, lesion regions, and an overall lesion subtype. Due to training logistics and available computing resources, the dataset was pre-cropped to smaller regions of interest.

TABLE I: Dataset Split

Split	Subtype 0	Subtype 1	Subtype 2
Train	62	106	84
Validation	9	15	12

The dataset comprises 252 training images, and 36 validation images. Since we used nnU-Net V2's 5-fold cross-validation, the validation set is use to evaluate the performance of our final model. Thus, nnU-Net V2 splits the training set during planning. No preprocessing was performed except for the preprocessing undertaken by nnU-Net V2 by default.

III. Methods

A. nnU-Net V2 Residual Encoder

nnU-Net V2 is a state-of-the-art deep learning framework designed for medical image segmentation. It provides automated pipelines for preprocessing, network configuration, and training. Building on the strengths of its predecessor, nnU-Net V2 incorporates a modular architecture, enhanced support for multi-task learning, and improved scalability, making it a robust choice for diverse medical datasets.

The proposed model utilized nnU-Net V2's Residual Encoder UNet architecture, specifically the nnU-Net ResEnc M preset in the 2D configuration. The 2D configuration was chosen mostly due to training logistics, with the 3D configuration

empirically being much slower to train. The architecture was then modified to add a classifier head in parallel to the existing UNetDecoder.

By default, ResEnc M uses dice and binary cross entropy loss with logits via an nnU-Net V2 function called DC_and_BCE_loss.

B. Classifier Head

The classification head is implemented as a sequential feed-forward neural network, designed to predict target class labels from encoded feature representations. The architecture begins with a convolutional feature extractor (CFE), using 2 layers of convolutions with kernel size of 2, and stride of 1. Both layers use ReLU activation and batch normalization. The output of the CFE is flattened and passed through 4 fully connected (dense) layers with batch normalization, ReLU activation, and a dropout probability of 0.3 except for the last layer which has a dropout of 0.2. Finally, the the 5th fully connected layer maps the feature representation to the number of target classes (3 in this case). A batch normalization layer is applied after this output, but no activation function is introduced at the output of the classifier head to align with the DC_and_BCE_loss function used for training (see the Discussion Section for details).

A detailed model architecture diagram with accompanying legend is presented in Figure 1.

For training, the proposed model uses a weighted sum of losses between both heads, with a weight of 1 for the segmentation head, and a weight of 3 for the classifier head. Due to the use of the 2D configuration, the proposed model outputs a single subtype prediction per slice of the CT scan. Since tumors are 3D, we ensemble these slice predictions to obtain the prediction for the full tumor.

IV. Training

Training was conducted with a modified nnUNetTrainer object to facilitate multi-task learning. Weighted sum of losses was used to train both segmentation and classifier heads. A weight of 3 was given to the classifier loss, due the segmentation head's tendency to dominate, especially early in the training. Other changes comprise primarily of classifier evaluators and metrics, and loggers. Training used the default stochastic gradient descent optimizer with a learning rate of 1e-2, and weight decay of 3e-5. Learning rate decay also used the default polynomial scheduler. Batch size was reduced to 16 due to VRAM limitations.

nnU-Net, by default, uses a 5-fold cross-validation scheme, using 5 unique splits of the same training data. To predict, these folds are independently trained and their outputs ensembled. An 80/20 split was used to produce each fold, yielding 5 splits of 201/202 training images and 51/50 validation images.

Training was conducted over 100 epochs as losses converged far ahead of the default of 1000 epochs. Each fold was independently trained on it's split. 5-fold loss graphs are included in figure 3.

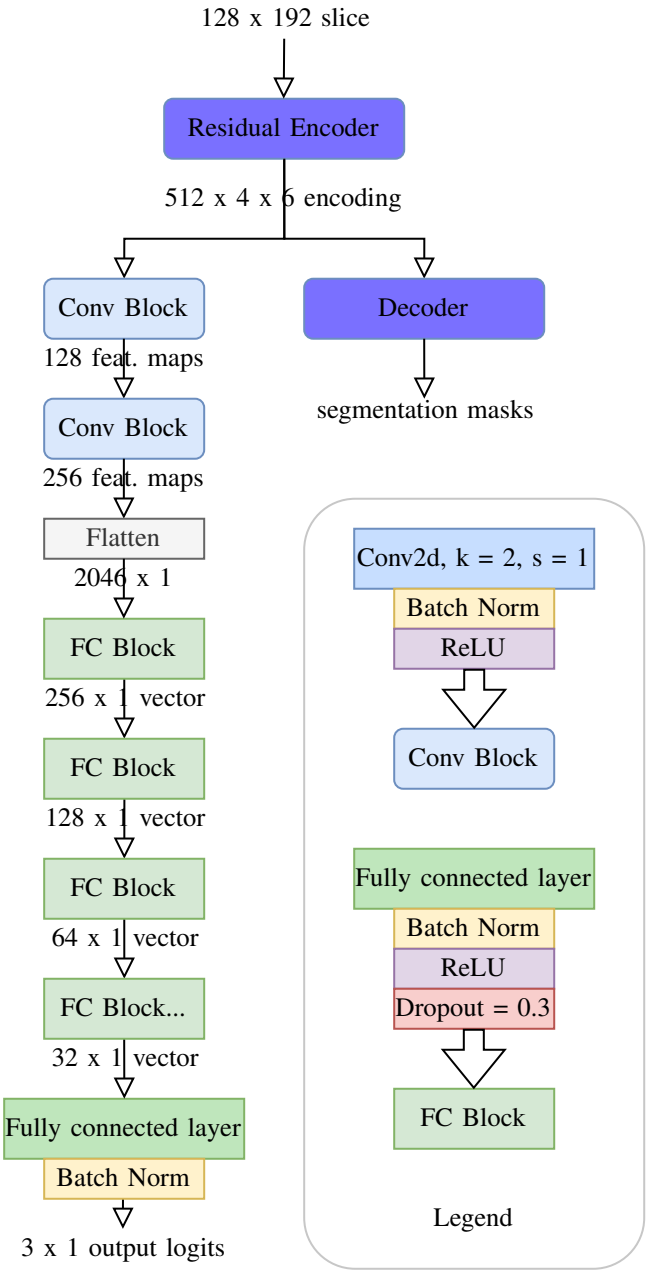


Fig. 1: Multi-task model architecture

V. Results

Metrics were chosen as per Metrics Reloaded [5].

A. Classification

After the 5-fold ensembling, the proposed model's classifier head achieved an accuracy of 91.67%, an F_β of 0.9105, average precision of 92.58%, and a brier score of 0.0611. Sensitivity and specificity per class were also evaluated.

The result is promising and yields average sensitivity and specificity of 90.4% and 91.4% respectively. Comparing to the results of [3], which had more lesion subtypes in their classification, the sensitivity is similar, although with a lower

TABLE II: Per class sensitivity and specificity

	Subtype 0	Subtype 1	Subtype 2
Sensitivity	77.8%	93.3%	100.0%
Specificity	96.3%	90.5%	87.5%

specificity by ˜6% between the models. This indicates performance comparable to human professionals, over the lesion types present in the dataset. The confusion matrix is also presented in figure 4. Of note is that the classifier tends to be most confident with subtype 2, with the highest false positive rate in subtype 1.

B. Segmentation

The segmentation head achieved a Dice of 45.36, and an IoU of 35.20%. The segmentation result is quite poor. In particular, the IoU score indicates the segmentation failed to properly label large portions of the ground truth. Part of this result may be due to the added strain from training a classifier at the same time or small dataset size. Additionally, architecture choices could have had a large impact, particularly the choice to use the 2D instead of 3D configuration for nnU-Net. An example of some segmentation results are shown in figure 2. Note that in both results, the segmentation result tended to predict floating regions outside the pancreas, along with feathering between slices. Both of these artifacts are likely related to using the 2D configuration, since each slice's segmentation prediction is not informed by the 3 dimensional structure vertically above and below it.

C. Ablation study

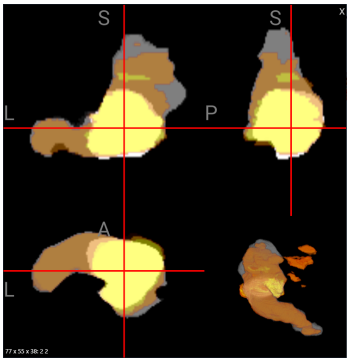
An ablation study was conducted on the classifier head, results indicate that the CFE makes a sizable contribution to the network, however, even a single dense layer provides a respectable 84.3% accuracy. Along the same trend, the rest of the results also show only mildly poorer results across all metrics. Removing batch normalization and dropout (which were investigated due to overfitting) both also give a mild performance decrease. Network depth did not make a major difference. These results indicate that model complexity is not the limiting factor in raising metric scores, and the problem likely lies with other aspects of the model, such as the encoder, or the training regimen, in particular, balancing segmentation and classification could be improved. Due to time constraints, ablation could not be run on the full 5-fold ensemble, and results for a single fold are shown. Of note is that ensembling made the largest difference to Dice, up to an increase of 15, but accuracy only increased a few percent (88.2% $\rightarrow$ 91.7%). All model variations (except 'No Batchnorm' and'No Dropout') are presented as progressive removals from the proposed architecture.

D. Discussion

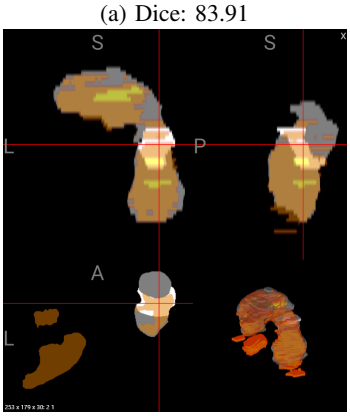
We chose to use DC_and_BCE_loss since empirically, it provided better balancing between our segmentation and classifier heads, which lead to better dice loss by ˜3 and accuracy by about 8% respectively (this, admittedly, needs more investigation into the base cause, using dice loss on what is ostensibly a single pixel is not proper).

One of the drawbacks of ensembling per-slice predictions was the unfairing weighting of slices. Those that did contain portions of tumor or pancreas contribute to the classification decision with the same weight as slices that do. Adding a fourth background class during data processing, so that slices could be weighed differently during ensembling, was attempted, but empirically performed worse than the proposed model. This can likely be attributed to an increase in difficulty to predict over an additional class.

It should be noted that different models during the ablation study sometimes performed better or worse than others for specific classes. This could be taken advantage of by implementing One-vs-Rest (OvR) or One-vs-All (OvA), and training separate several classifier heads instead of one head outputting



(a) Dice: 83.91



(b) Dice: 10.95

Fig. 2: Strong vs weak segmentation results.
Note: predicted segmentation is in orange

TABLE III: Ablation Study

Model Variation	Accuracy	Brier Score	Dice
Proposed Model	88.2%	0.061%	30.9
No Batchnorm	86.2%	0.064%	30.5
No Dropout	84.3%	0.078%	30.2
No CFE	82.3%	0.068%	31.4
3 dense layers	85.3%	0.056%	30.7
1 dense layer	84.3%	0.083%	29.7

a 3-long softmax output. additionally, as we saw in the ablation study, classifier architecture is not a limiting factor. To mitigate this, one possible alternative is to first train the model only for segmentation, then freeze these weights before training only the classifier head.

The choice of the 2D configuration instead of 3D also needs further investigation. During training, the 3D configuration ran significantly slower, 3-5 minutes per epoch versus 20 seconds for 2D. This meant that a full 5-fold training at 100 epochs would have taken over 30 hours when including inference and evaluation. This would have heavily restricted how many variations of architecture we could have tested.

VI. Code Release

Model implementation and reproduction instructions can be found here:

https://github.com/davidguo123456/pancreas-cancer-segmentation

VII. Conclusion

In this study, we extended the capabilities of nnU-Net V2 to perform simultaneous segmentation of pancreatic tumors and classification of their subtypes using a multi-task architecture. The proposed model integrates a shared encoder with dual decoder heads for segmentation and classification, enabling learning from a shared feature space. By leveraging nnU-Net's automated pipelines and performing 5-fold cross-validation, we achieved promising classification performance, with an accuracy of 91.67% and average sensitivity and specificity of 90.4% and 91.4%, respectively. These results demonstrate the potential for deep learning to provide non-invasive and accurate subtype classifications, comparable to human performance.

However, the segmentation head's performance, with a Dice score of 45.36 and an IoU of 35.20%, highlights the challenges of achieving high-quality segmentation in the presence of small datasets and architectural constraints. The choice of a 2D configuration over 3D, while dictated by computational limitations, may have contributed to suboptimal segmentation outcomes.

Our ablation study further highlights the importance of fully connected layers, batch normalization, and ensembling for robust classification performance. Attempts to improve prediction weighting during ensembling through the introduction of a fourth background class showed limited success, suggesting areas for further exploration.

Future work could focus on addressing these limitations, including exploring 3D configurations, expanding the dataset, exploring alternative training regimensm, and refining ensembling techniques to improve segmentation quality. Additionally, evaluating the model on external datasets will further validate its robustness and ability to generalize.

References

[1] A. McGuigan, P. Kelly, R. C. Turkington, C. Jones, H. G. Coleman, and R. S. McCain, "Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes," *World Journal of Gastroenterology*, vol. 24, no. 43, p. 4846–4861, Nov 2018.

[2] C. Torres and P. J. Grippo, "Pancreatic cancer subtypes: A roadmap for precision medicine," *Annals of Medicine*, vol. 50, no. 4, p. 277–287, Mar 2018.

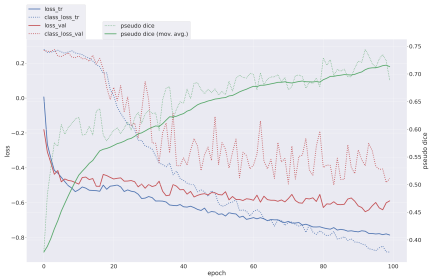
[3] K. Cao, Y. Xia, J. Yao, X. Han, L. Lambert, T. Zhang, W. Tang, G. Jin, H. Jiang, X. Fang, and et al., "Large-scale pancreatic cancer detection via non-contrast ct and deep learning," *Nature Medicine*, vol. 29, no. 12, p. 3033–3043, Nov 2023.

[4] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, p. 203–211, Dec 2020.

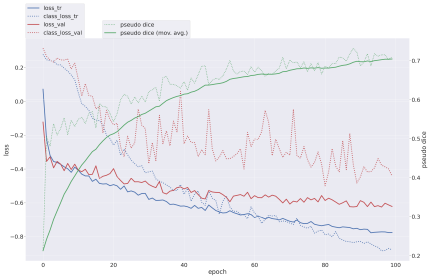
[5] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, and et al., "Metrics reloaded: Recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, p. 195–212, Feb 2024.

APPENDIX

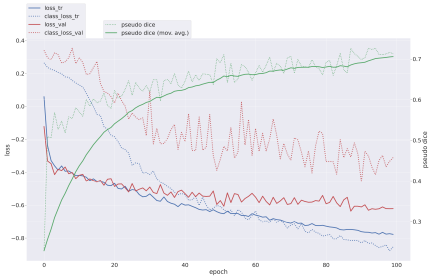
Fig. 3: Loss graphs for all folds



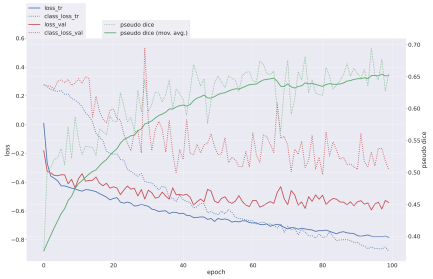
(a) Fold 0



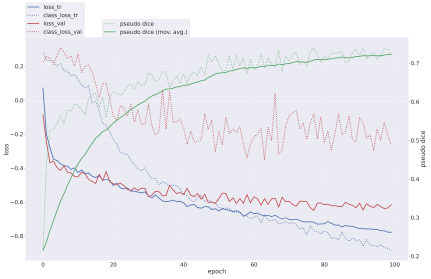
(b) Fold 1



(c) Fold 2



(d) Fold 3



(e) Fold 4

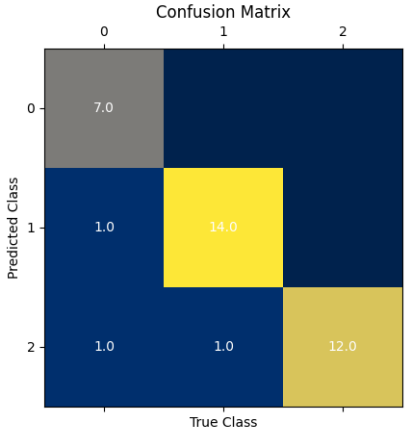


Fig. 4: Classifier head confusion matrix