Accurate and Feasible Deep Learning Based Semi-automatic Segmentation in CT for Radiomics Analysis in Pancreatic Neuroendocrine Neoplasms - *Supplementary Materials*

I. CT ACQUISITION AND RECONSTRUCTION

Patients from Center 1 were examined by performing contrast-enhanced abdominal CT scans on a TOSHIBA Aquilion scanner (TOSHIBA Medical, Tokyo, Japan). The reconstructed image matrix was 512 ×512, and the slice thickness was 1 mm. Patients from Center 2 were scanned using contrast-enhanced abdominal CT on a Discovery CT750 HD scanner (GE Healthcare, Milwaukee, USA). The image matrix was 512 ×512, and the slice thickness was 0.9mm - 2 mm.

II. PATHOLOGICAL GRADING OF PNENS

In Dataset 1, the pathological grade of patients was re-assessed separately and determined by consensus of two pathologists (L.X. and Y.L.) with 11 years of experience in abdominal cancer diagnosis. It should be noted that these two pathologists were blinded to each other's presence and the results of the previous diagnosis before assessing the pathological grade. Different modalities of specimen collection were first performed for patients with different treatments. For patients who underwent surgical resection, 3 specimens from different areas of the tumor were taken; biopsies guided by ultrasound were taken for patients who could not undergo surgical resection. Immunohistochemical staining of somatostatin receptor was then performed on all of the specimens and the mitotic count and Ki67 index were counted for pathological grading.

III. DIAGNOSTIC CRITERIA OF PROGNOSIS AFTER CURATIVE SURGERY

For the patients in Dataset 2, the follow-up was conducted from the third month after surgery until May 2019 (overall median follow-up, 1320 days). At least one biannual medical imaging examination (ultrasound/CT/MRI) was carried out during the first year and henceforth on either a biannual or annual basis depending on the tumor grading (G1, every year; G2/3, every 6 months). Positron Emission Computed Tomography (PET-CT) with ^{68}Ga-labelled somatostatin analogue and ^{18}F-fluorodeoxyglucose were reserved for doubtful patients of disease recurrence. Postoperative recurrence was determined by agreement between two abdominal radiologists (C.S. with 3 years experience and Y.L. with 11 years experience) after a joint discussion combining the patient's history, pathological findings, and imaging examination. Similar to studies [1, 2], the date of recurrence was defined as the time when recurrence was discovered based on cross-sectional imaging during 3-year follow-up (CT/MRI).

IV. FEATURE EXTRACTION AND SELECTION

As suggested by Vallières et al. [3] , all tumor volumes were first resampled to the same voxel size (0.675mm × 0.675mm × 0.8 mm) using three-dimensional cubic interpolation on MATLAB software (64-bit version R2019a; The MathWorks, Natick, MA). Next, by using our own feature kit based on MATLAB, 167 radiomics features were extracted including 24 shape features, 37 intensity features, and 106 texture features, such as gray-level co-occurrence matrix, gray-level run-length matrix, gray-level size zone matrix, spatial gray-level dependence matrix, neighborhood gray-tone difference matrix, and neighborhood gray-level difference statistics. Before calculating the texture features, the uniform quantization algorithm was chosen to quantify the entire intensity range of the tumor region to gray levels of 1 - 64.

With the training dataset, the best feature subset was obtained by using the least absolute shrinkage and selection operator (LASSO) algorithms [4]. The selected radiomics features were then normalized as z-scores to train the model. With the testing dataset, the features were normalized in the same way except that the mean and the standard deviation were calculated within the training dataset.

V. THE DETERMINATION OF SAMPLE SIZE

As reported in studies [5, 6], the determination of sample size was performed in our study.

For the internal cross-validation (CV) of the pathological grading prediction study, we investigated whether the sample size in our study was sufficient to detect an statistical difference based on the following conditions by using MedCalc Statistical Software version 15.8 (MedCalc Software bvba, Ostend, Belgium, https://www.medcalc.org/manual/sampling_ROC1.php): statistical power, 80%; a two-tailed significance level, 0.05; the true AUC values of the model with semi-automatic segmentation for pathological grading prediction; the null hypothesis of AUC=0.5; ratio of classes, the real ratios in our study, for example 68 (G1/2) / 12 (G3) in the internal dataset of pathohistologic grading.

For recurrence prediction study, we also estimated the required the sample size based on following input: statistical power, 80%; a two-tailed significance level, 0.05; the true AUC values of model with semi-automatic segmentation for recurrence prediction; the null hypothesis of AUC = 0.5; ratio of classes, the real number of positive and negative cases in recurrence prediction study.

TABLE S1

COMPARISON OF SEGMENTATION PERFORMANCE BETWEEN OUR PROPOSED METHOD AND OTHER METHODS.

Study	Tumor type	Algorithm	Segmentation method	DSC [a] Mean ± SD [b] (%)
Zhou et al.[7]	Pancreatic Cyst	2D without Deep supervision	Automatic	60.5 ± 31.4
		2D with Deep supervision	Automatic	63.4 ± 27.7
		2D with reference standard Cyst B-box	Semi-automatic	77.9 ± 12.8
Guo et al.[8]	Pancreatic tumors	3D Deep LOGISMOS	Semi-automatic	83.2 ± 7.8
Zhou et al.[9]	PDAC [c]	3D-UNet-multi-phase-HPN	Automatic	57.1 ± 24.8
		3D-ResDSN-multi-HPN	Automatic	63.9 ± 22.7
Ours	**pNENs** [d]	**2D-UNet (Internal cross-validation)**	**Semi-automatic**	**79.6 ± 13.0**
		2D-UNet (External validation)	**Semi-automatic**	**81.8 ± 8.4**

[a] Dice similarity coefficient; [b] Standard deviation; [c] Pancreatic Ductal Adenocarcinom; [d] Pancreatic Neuroendocrine Neoplasms**.**

TABLE S2

CORRELATION ANALYSIS RESULTS OF THE SAME 4 FEATURES (AS SHOWN IN FIGURE S2) BETWEEN MODELS WITH SEMI-AUTOMATIC AND MANUAL SEGMENTATION AFTER FEATURE SELECTION.

	Variable	r	*P* value
Intensity-based features	Max	0.85	<0.001
	90^{th} percentile	0.97	<0.001
SGLDM[a]-based features	Large Dependence High Gray Level Emphasis_ LDHGLE	0.76	<0.001
NGLDS[b]-based features	rMeanAD	0.13	0.27

[a] Spatial gray-level dependence matrix; [b] Neighborhood gray-level difference statistics.

TABLE S3

COMPARISON OF PATHOLOGICAL GRADING PREDICTION PERFORMANCE BETWEEN OUR MODELS (80 PATIENTS FROM CENTER 1) AND OTHER STUDIES.

Classifiers	Segmentation method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC [e] (95% CI [f])
LR [a]	**Semi-automatic (ours)**	76	75	83	0.76 (0.65 - 0.85)
	Manual (ours)	85	88	67	0.72 (0.61 - 0.82)
	Manual (Luo et al.[10])	73	75	72	0.72 (0.62 - 0.81)
SVM [b]	**Semi-automatic (ours)**	73	71	83	0.71 (0.59 - 0.80)
	Manual	86	88	75	0.80 (0.69 - 0.88)
	Manual (Luo et al.[10])	71	88	68	0.80 (0.71 - 0.88)
RF [c]	**Semi-automatic (ours)**	68	65	83	0.73 (0.62 - 0.82)
	Manual	81	82	75	0.79 (0.68 - 0.87)
	Manual (Luo et al.[10])	70	75	69	0.72 (0.62 - 0.81)

[a] Logistic regression; [b] Support vector machine; [c] Random forest; [e] Area under the receiver; [f] Confidence interval.

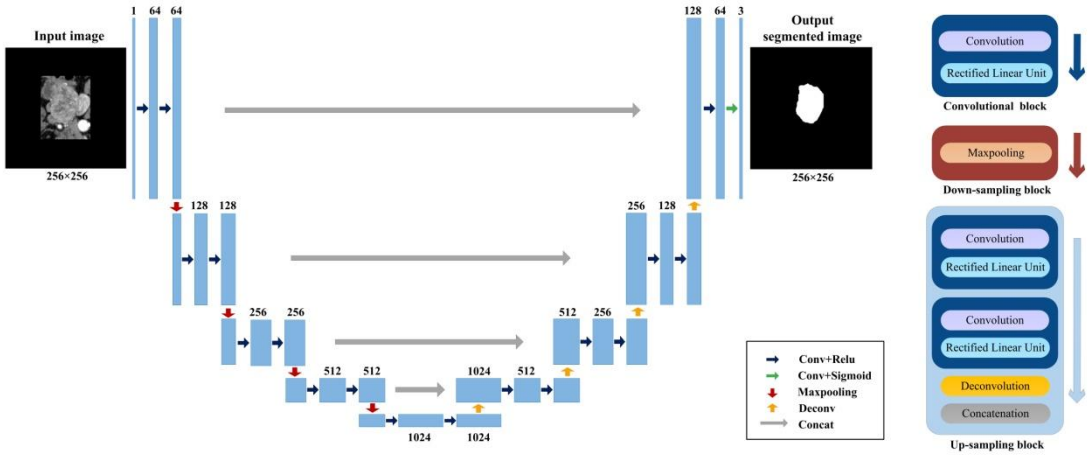

Fig. S1. Architecture of the U-Net network. Each blue rectangle corresponds to a set of feature maps. The number above each blue rectangle represents the number of filters. Different colored arrows indicate different operations. The details of the convolution block, the down-sampling block and the up-sampling block are shown on the right.

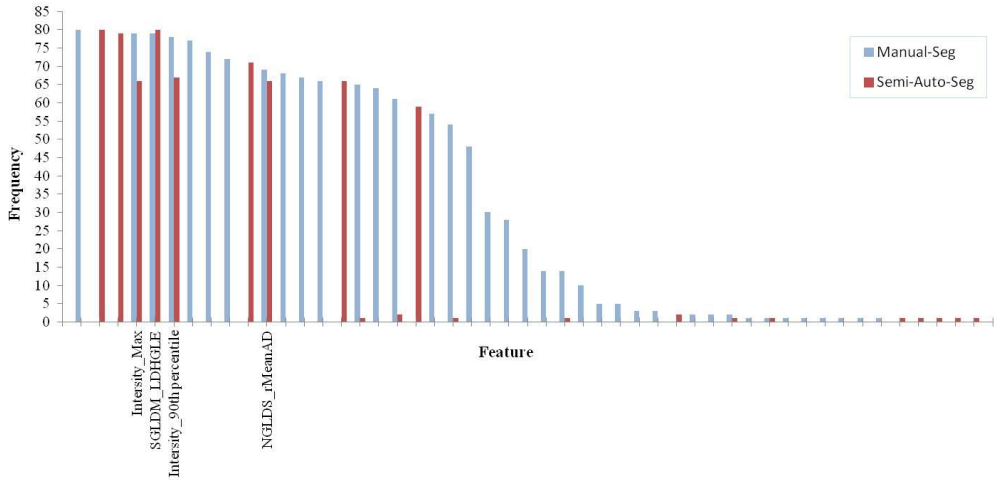

Fig. S2. The frequency of the features selected in cross validation of the models with semi-automatic segmentation (red bar) and the models with manual segmentation (blue bar) respectively. The vertical axis indicates the frequency. The horizontal axis indicates different features. Note that 4 features appear frequently (>58 in leave-one-out cross validation) in both models with semi-automatic segmentation and manual segmentation. SGLDM, spatial gray-level dependence matrix; LDHGLE, large dependence high gray level emphasis; NGLDS, neighborhood gray-level difference statistics.

REFERENCES

[1] A. Kaltenborn, S. Matzke, M. Kleine, T. Krech, W. Ramackers, F. W. R. Vondran, et al., "Prediction of survival and tumor recurrence in patients undergoing surgery for pancreatic neuroendocrine neoplasms," Journal of Surgical Oncology, vol. 113, no. 2, pp. 194-202, 2016.

[2] S. Partelli, A. A. Javed, V. Andreasi, J. He, F. Muffatti, M. J. Weiss, et al., "The number of positive nodes accurately predicts recurrence after pancreaticoduodenectomy for nonfunctioning neuroendocrine neoplasms," European Journal of Surgical Oncology, vol. 44, no. 6, pp. 778-783, 2018.

[3] M. Vallieres, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," Phys Med Biol, vol. 60, no. 14, pp. 5471-5496, 2015.

[4] R. Tibshirani, "Regression shrinkage and selection via the Lasso," J Roy Stat Soc B Met, vol. 58, no. 1, pp. 267-288, 1996.

[5] J. E. Park, H. S. Kim, S. Y. Park, S. J. Nam, S.-M. Chun, Y. Jo, et al., "Prediction of Core Signaling Pathway by Using Diffusion- and Perfusion-based MRI Radiomics and Next-generation Sequencing in Isocitrate Dehydrogenase Wild-type Glioblastoma," Radiology, vol. 294, no. 2, pp. 388-397, 2020.

[6] L. Dercle, M. Fronheiser, L. Lu, S. Du, W. Hayes, D. K. Leung, et al., "Identification of Non–Small Cell Lung Cancer Sensitive to Systemic Cancer Therapies Using Radiomics," Clinical Cancer Research, vol. 26, no. 9, pp. 2151-2162, 2020.

[7] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille, "Deep supervision for pancreatic cyst segmentation in abdominal

CT scans," Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2017, pp. 222-230.

[8] Z. Guo, L. Zhang, L. Lu, M. Bagheri, R. M. Summers, M. Sonka, et al., "Deep LOGISMOS: Deep learning graph-based 3D segmentation of pancreatic tumors on CT scans," Proc. IEEE 15th International Symposium on Biomedical Imaging, IEEE, Washington, DC, 2018, pp. 1230-1233.

[9] Y. Zhou, Y. Li, Z. Zhang, Y. Wang, A. Wang, E. K. Fishman, et al., "Hyper-Pairing Network for Multi-phase Pancreatic Ductal Adenocarcinoma Segmentation," Proc. International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, Cham, 10 Oct, 2019, pp. 155-163.

[10] Y. Luo, X. Chen, J. Chen, C. Song, J. Shen, H. Xiao, et al., "Preoperative Prediction of Pancreatic Neuroendocrine Neoplasms Grading Based on Enhanced Computed Tomography Imaging: Validation of Deep Learning with a Convolutional Neural Network," Neuroendocrinology, vol. 110, no. 5, pp. 338-350, 2020.