Preoperative Prediction of Pancreatic Neuroendocrine Neoplasms Grading Based on Enhanced Computed Tomography Imaging: Validation of Deep Learning with a Convolutional Neural Network

Yanji Luo^a Xin Chen^b Jie Chen^c Chenyu Song^a Jingxian Shen^d
Huanhui Xiao^b Minhu Chen^c Zi-Ping Li^a Bingsheng Huang^b Shi-Ting Feng^a

^aDepartment of Radiology, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou, China;

^bSchool of Biomedical Engineering, Health Science Center, Shenzhen University, Block A2, Xili Campus of

Shenzhen University, Shenzhen, China; ^cDepartment of Gastroenterology, The First Affiliated Hospital,

Sun Yat-Sen University, Guangzhou, China; ^dDepartment of Radiology, Sun Yat-Sen University Cancer Center, Guangzhou, China

Keywords

Pancreatic neoplasms · Neuroendocrine tumors · Deep learning · Tomography, spiral computed · Neoplasm grading

Abstract

Introduction: The pathological grading of pancreatic neuroendocrine neoplasms (pNENs) is an independent predictor of survival and indicator for treatment. Deep learning (DL) with a convolutional neural network (CNN) may improve the preoperative prediction of pNEN grading. **Methods:** Ninety-three pNEN patients with preoperative contrast-enhanced computed tomography (CECT) from Hospital I were retrospectively enrolled. A CNN-based DL algorithm was applied to the CECT images to obtain 3 models (arterial, venous, and arterial/venous models), the performances of which were evaluated via an eightfold cross-validation technique. The CECT images of the optimal phase were used for comparing

the DL and traditional machine learning (TML) models in predicting the pathological grading of pNENs. The performance of radiologists by using qualitative and quantitative computed tomography findings was also evaluated. The best DL model from the eightfold cross-validation was evaluated on an independent testing set of 19 patients from Hospital II who were scanned on a different scanner. The Kaplan-Meier (KM) analysis was employed for survival analysis. **Results:** The area under the curve (AUC; 0.81) of arterial phase in validation set was significantly higher than those of venous (AUC 0.57, $p = 0.03$) and arterial/venous phase (AUC 0.70, $p = 0.03$) in predicting the pathological grading of pNENs. Compared with the TML models, the DL model gave a higher (although insignificantly) AUC. The highest OR was achieved for the p ratio <0.9 , the AUC and accuracy for diagnosing G3 pNENs were 0.80 and 79.1% respectively. The DL algorithm achieved an AUC of 0.82 and an accuracy of 88.1% for the independent testing set. The KM analysis showed a statistical significant difference between the predicted G1/2

Y.L., X.C., and J.C. contributed equally to this work.

Bingsheng Huang
School of Biomedical Engineering, Health Science Center
Shenzhen University, Block A2, Xili Campus of Shenzhen University
1066 Xueyuan Avenue, Shenzhen (China)
E-Mail huangb@szu.edu.cn

Shi-Ting Feng
Department of Radiology, The First Affiliated Hospital
Sun Yat-Sen University, 58th, The Second Zhongshan Road
Guangzhou, Guangdong (China)
E-Mail fengsht@mail.sysu.edu.cn

and G3 groups in the progression-free survival ($p = 0.001$) and overall survival ($p < 0.001$). **Conclusion:** The CNN-based DL method showed a relatively robust performance in predicting pathological grading of pNENs from CECT images.

© 2019 S. Karger AG, Basel

Introduction

Pancreatic neuroendocrine neoplasms (pNENs) are the second most common pancreatic malignancy emanating from the neuroendocrine cells of the pancreas [1, 2]. According to the 2010 WHO classification system, pNENs were classified into 3 grades based on their mitotic count and Ki-67 index: low-grade as G1 (mitotic count <2 per 10 high power fields (HPFs) and/or a Ki67 index of $<3\%$), intermediate-grade as G2 (mitotic count 2–20 per 10 HPF and/or a Ki67 index of 3–20%) and high-grade as G3 (mitotic count >20 per 10 HPF and/or a Ki67 index of $>20\%$). In the 2010 WHO classification system, G3 pNENs have been classified as poorly differentiated neuroendocrine carcinomas (pNECs) [3, 4]. However, over the years, a number of reports have suggested that G3 pNECs based on 2010 WHO classification are more heterogeneous than expected. Recently, the 2017 WHO classification system has identified a subset of G3 well-differentiated pancreatic neuroendocrine tumors (pNETs) where both G3 pNETs and pNECs are classified as G3 pNENs. Based on the mitotic count and Ki67 index, the grading of G1/2 and G3 pNENs was similar when comparing the 2017 WHO classification system with the 2010 WHO classification system. The only difference is that the 2017 WHO classification system identified a subset of G3 pNETs from the former G3 tumors. Treatment response data suggest that both G3 pNETs and pNECs responded to both platinum and alkylating agents [5]. Early investigation of G3 pNETs has demonstrated survival times to be shorter than those for G1 or G2 pNETs, but longer than what is typically described for pNECs. The 5-year survival rates ranged from 60 to 100% for G1/2 and 16 to 29% for G3 pNENs [5, 6].

Under most circumstances, the pathological features of pNEN tumors were determined after the lesion is resected. Tumor resections may be associated with postoperative complications [7, 8]. Computed tomography (CT) plays a crucial role in the diagnosis of pNENs. A previous study revealed that there are significant differences in several qualitative and quantitative CT image features between the G1/2 and G3 pNENs [9, 10]. But such evaluations are subjective, have poor repeatability, and/or are time-consuming. To this end, radiomics strategies involving a

high-throughput extraction of the quantitative parameters from the CT images have been developed to grade the pNENs [11, 12]. However, a radiomics analysis requires manually contouring the tumor lesions as the regions of interest (ROI), during which the inter-observer variability cannot be avoided [13]. In contrast, deep learning (DL) with a convolutional neural network (CNN) does not require drawing the ROIs along the tumor margin, and the feature extraction and selection can be performed automatically, demonstrating a higher performance during clinical applications [14–18]. Although 2 literatures have reported about predicting pathological grading of pNENs by using the radiomics method (Table 1) [11, 12], the application of DL with CT imaging to predict pathological grading of pNENs has not yet been reported.

We understand that CT cannot replace histopathology assessment when determining the grade of pNENs and subsequently the choice of treatment based on CT features alone. However, preoperative information of G3 pNENs via CT imaging may be helpful in estimating the tumor aggressiveness and provide a reliable basis for treatment plan to be made preoperatively. If there are CT features suggesting a G3 pNENs, liver MRI, including diffusion-weighted and hepatobiliary specific contrast agent (Gadolinium-ethoxybenzyl-diethylenetriamine pentaacetic acid, Gd-EOB-DTPA) enhanced MRI, may be helpful in discerning liver metastases. As such, identifying the more aggressive pNENs preoperatively could stratify a tailored treatment plan for the patients, in which the G1/2 group could receive parenchyma-sparing pancreatic resection, while the G3 group might undergo comprehensive treatment strategies including radical surgical resection and systematic chemotherapy to improve long-term prognosis [19, 20].

In this study, we have trained a DL model to predict the pathological grading of pNENs from contrast-enhanced CT (CECT) images via a noninvasive approach. We compared the performance of the DL model with those of a traditional machine learning (TML) model where the radiologists validated its robustness on an independent testing set (Fig. 1).

Materials and Methods

Patient Selection

The protocol for this retrospective study was approved by the Institutional Review Board of Sun Yat-Sen University. The procedures conducted in the study were adherent to the principles of the Declaration of Helsinki. A written informed consent was obtained from each patient. The patients were selected from the pathology files attending to our institutions between January 2010 and June

Fig. 1. Flowchart of the study design. CECT, contrast-enhanced computed tomography; pNEN, pancreatic neuroendocrine neoplasm; DL, deep learning; CNN, convolutional neural network; TML, traditional machine learning.

Table 1. Summarization of the relevant studies on predicting the pathological grading of pNENs from CT images by using radiomics methods

	Literature 1	Literature 2
Sample size	137 patients from 2 institutions (86 and 51) The data from institution 2 were used as an independent testing dataset	101 patients from one institution
CT images used in the study	Arterial phase	Portal venous phase
Radiomics method	A combined nomogram model incorporating the radiomics signature without/with clinical factors	CT texture analysis
Classification	Differentiating G2/3 from G1 tumors	Differentiating G2/3 from G1 tumors
Results	Radiomics signature model: training set, AUC 0.870; testing set, AUC 0.862 Combining radiomics signature with clinical stage: training set: AUC 0.907; testing set: AUC 0.891	Entropy and all filters model: accuracy, 65% Combining entropy and the CT features: accuracy, 79.3%
Prognosis analysis	The OS of predicted G1 and G2/3 groups was significantly different ($p = 0.002$)	The PFS of G1 and G2/3 groups was significantly different ($p < 0.001$)

OS, overall survival; PFS, progression-free survival; CT, computed tomography; pNENs, pancreatic neuroendocrine neoplasms; AUC, area under the curve.

2017. The CECT data were available for all patients, and none of the patients received any treatment at the time of imaging. Finally, 93 patients from The First Affiliated Hospital of Sun Yat-Sen University (Hospital I) and 19 patients from The Cancer Center of Sun Yat-Sen University (Hospital II) with pathologically confirmed pNENs were enrolled.

Histological Analysis

The specimens were collected and re-evaluated by 2 (X.L. and L.Y.) pathologists with 10 years of experience in diagnosing abdominal tumor. Immunohistochemical staining of somatostatin receptor were performed on all of the specimens. The pathological grading was made based on the WHO 2017 classification system [4]. The well differentiated G3 pNETs and poorly differentiated

pNECs were identified using the 2017 WHO classification system. The results showed that there were 4 and 2 pNETs from Hospital I (G3 pNETs: $n = 4$, pNECs: $n = 13$) and Hospital II (G3 pNETs: $n = 2$, pNECs: $n = 4$) respectively. The 2 pathologists agreed on the final pathological grading of the pNENs.

For patients with surgical resection of the pancreatic tumor, 3 specimens from different regions of the tumor were taken for histological analysis. For those with distant or diffused metastasis where curative resection was not an option, biopsy guided by ultrasound, rather than surgical resection were performed on the primary pancreatic lesion where 1 (for tumor size ≤ 2 cm) or 3 (for tumor size > 2 cm) samples were obtained from different areas of the tumor. All the biopsy specimens were used in the final pathological analysis. It is important to note that both the mitotic count

Fig. 2. Example of data preprocessing. **a** Original CECT image (512×512). **b** Tumor ROI. **c** The tumor ROI is centered in the 3D matrix ($280 \times 280 \times z$; z is the number of image slices of the tumor ROI). **d** The 3D matrix is split into several image blocks with each red region indicating one block.

and Ki67 index can be heterogeneous in the same pancreatic neuroendocrine lesions. Thus, WHO recommends counting 500–2,000 neoplastic cells and manual counting of camera-captured or printed images of a “hot-spot” is considered to be the most practical and reproducible approach for calculating the mitotic count and Ki67 index [5]. The above measures were applied to avoid imprecise pathological grading due to tumor heterogeneity.

CT Image Acquisition

The CT images of the 93 (G1/2: $n = 76$, G3: $n = 17$) patients from Hospital I were captured using a 64-slice spiral CT scanner (Aquilion 64, Canon Medical Systems) with the following parameters: slice thickness, 0.5 mm; slice interval, 0.5 mm; tube current, 200 mAs; tube voltage, 120 kVp. Following the pre-contrast imaging, an iodinated contrast (Ultravist 300, Bayer Schering, Berlin) was administered intravenously at a rate of 3 mL/s via a high-pressure syringe. The arterial and portal venous phases were obtained at 35 and 65 s post contrast injection respectively. All contrast injections were followed by a saline chaser bolus (40 mL) at the same rate.

The CT images of the 19 (G1/2: $n = 13$, G3: $n = 6$) patients from Hospital II were captured using a 128-slice spiral CT system (Discovery CT750 HD, GE System, Milwaukee, WI, USA) with the following parameters: slice thickness, 2 mm; slice interval, 1 mm; tube voltage, 100–140 kVp (12 patients at 120 kVp, 4 patients at 100 kVp, and 3 patients at 140 kVp); tube current, automatic tube current modulation (maximum 450 mAs). The CECT was performed after injecting a non-ionic iodinated contrast agent (Ultravist 300, Bayer Schering, Berlin) intravenously into the antecubital vein at a rate of 3 mL/s via a high-pressure syringe. The arterial and portal venous phases were obtained at 10 and 35 s delays, respectively, after the aortic opacification reached 100 HU, the average scan delay after contrast injection was 36 s (range 30–42 s) for the arterial phase and 66 s (range 58–70 s) for the portal venous phase.

DL for Pathological Grading of pNENs

Data Preprocessing

Using the open source software ITK-SNAP version 3.4 (<http://www.itk-snap.org>), a radiologist (S.C.) drew a rectangular ROI

Fig. 3. 3D CNN network structure. The settings of the filter number in the 3D CNN are the same as in Res50 [20]. For the first Conv3D layer, the filter size is (7,7,7) and the stride is (2,2,1). For the Pooling layer, the pool size is (3,3,3), and the stride is (2,2,1). For the IdentityBlock, the stride is (1,1,1) and the filter sizes are (1,1,1), (3,3,3), and (1,1,1) in the Conv3D layers from left to right.

For the Conv3D layers connected to the input layer directly in the ConvBlock, the filter size is (2,2,2), and the stride is (2,2,2). For the other 2 Conv3D layers in the ConvBlock, the stride is (1,1,1) and the filter sizes are (3,3,3) and (1,1,1) from left to right. Conv3D, 3D convolution layer; ReLU, rectifier linear unit layer; Pooling, 3D max pooling layer; GlobalPooling, 3D global average pooling layer.

containing the entire tumor lesion for each patient. The size of the ROI was different for different patients. For the CNN, the input images should have the same size; however, it was difficult to take the entire image as the network input because of the limited graphics processing unit memory. Therefore, we reconstructed the images of each tumor into several $280 \times 280 \times 16$ image blocks (Fig. 2) using the python image processing tool Opencv-Python 3.3.0.10 (<https://pypi.org/project/opencv-python/>). First, a blank $280 \times 280 \times z$ matrix was defined for each tumor lesion, where z represents the image slice number of the tumor lesion. The corresponding tumor lesion was then centered in the space defined by this matrix. The three-dimensional (3D) matrix was then sampled along the z -axis into several image blocks, with each block containing 16 slices. The remaining images, probably fewer than 16 slices, were treated as one single block. Each block was input into the CNN model and was labeled based on the pathological grade of the entire tumor (G1/2 as 0, and G3 as 1).

For data augmentation, an overlap of 75% was set between every 2 adjacent image blocks during the image preprocessing stage. Image rotation, contrast enhancement, and addition of Gaussian noise were done for data augmentation for CNN training.

Network Structure

Inspired by the deep residual learning network Res50 [21], a 3D CNN was applied in our study. The 3D CNN was composed of 1 convolution layer with 1 rectifier linear unit layer, a max pooling layer, 12 IdentityBlock, 4 ConvBlock, 1 global average pooling layer, and 1 fully connected layer with the sigmoid activation function (Fig. 3).

Network Training and Validation with DL Method

The data from Hospital I were used for the network training and validation. We trained 3 DL models – arterial model, venous model, and arterial/venous model – based on the arterial phase, venous phase, and a combination of the arterial and venous phases respectively. Each network was trained using the stochastic gradient descent optimizer for 300,000 iterations with a fixed learning rate of 1×10^{-6} . The DL toolbox Keras 2.1.1 (<https://github.com/keras-team/keras>) was used for the DL training and testing on a computer with a GeForce GTX 1080 graphics processing unit, a Core Intel Xeon E5-2650 V3 CPU clocked at 2.30 GHz, and a 64 GB random access memory.

As the images of a tumor lesion were divided into several image blocks, we obtained the patient-based testing result as follows. First, we denoted patient i ($0 < i < N+1$; N is the patient number) as $X_i = \{x_{ij}, j \in [1, \dots, M_i]\}$, where x_{ij} is the j^{th} image block, and M_i is the number of image blocks. With the learned 3D CNN, a probability p_{ij} for an image block x_{ij} was provided. The label of patient i was calculated using the following equation:

$$p_i = \frac{1}{M_i} \sum_{j=1}^{M_i} p_{ij}. \quad (1)$$

An eightfold cross-validation technique was applied to each of the models, and the performance of each model was assessed using the receiver operating characteristic (ROC) curve analysis. From the ROC curve, the area under the curve (AUC), accuracy, sensitivity, specificity, and the corresponding 95% CI were calculated.

Table 2. Accuracy, sensitivity, specificity, and AUC of grading pNENs into G1/G2 and G3 by using deep learning (93 patients from Hospital I)

	Accuracy, %	Sensitivity, %	Specificity, %	AUC (95% CI)	<i>p</i> value
Arterial model	73.1	82.4	71.1	0.81 (0.71–0.88)	–
Venous model	63.4	58.8	64.5	0.57 (0.46–0.67)	0.03
Arterial/venous model	69.9	64.7	71.1	0.70 (0.593–0.79)	0.03

The 95% CI is shown in the brackets. *p* value indicates the significance level of comparison of AUC with that of arterial model. pNENs, pancreatic neuroendocrine neoplasms; AUC, area under the curve.

Fig. 4. ROC curves of arterial, venous, and arterial/venous models for grading pNENs into G1/2 and G3 with DL method. AUC, area under the curve.

To compare the AUCs of the 3 models, a DeLong test was performed using the MedCalc software 18.5 (<https://www.medcalc.org/>). The 2-sided analysis with a *p* value <0.05 was considered to be statistical significant.

The input was the contrast-enhanced CT images inside a rectangular ROI containing the whole primary lesion. The output was the pathological grading (0 as G1/2, and 1 as G3) according to the biopsies representing the pNENs disease.

Comparison of DL with TML

To compare the DL and TML models, the data with the optimal phase of Hospital I (given in Section 4.3) were used in the TML model to predict the pathological grading of the pNENs. Three TML classifiers, namely, logistic regression (LR), random forest (RF), and support vector machine (SVM) were studied.

The implementation of a TML method requires the ground truth of the tumor lesions, which were manually drawn by 2 radiologists,

who knew the presence of pNENs but were blinded to the grade. The first radiologist, with 10 years of experience (L.Y.), performed all the contouring along the margin of the tumor in the thin axial CT images of the optimal phase by using ITK-SNAP version 3.4 (<http://www.itk-snap.org>). The manual segmentation was then examined by another radiologist with 20 years of experience (F.S.). Discrepancy was resolved by consensus after joint re-evaluation of the images.

The training set was divided into 2 groups (Group 1: patients with G1/2 tumors, Group 2: patients with G3 tumors) and >8,000 radiomics features were extracted by using the radiomics toolbox (<https://github.com/mvallieres/radiomics>), including the texture-based features such as gray level co-occurrence matrix, gray level run-length matrix, gray level size zone matrix, neighborhood gray-tone difference matrix, the nontexture features such as SUV metrics, AUC-CSH, percent inactive, size, solidity and volume, and the wavelet features, and so on. Two-sample *t* test or Mann-Whitney U test was performed, and the features which showed significant difference between these 2 groups were selected.

The features selected in the training set were then used in the validation set. The TML classifiers (LR, RF, and SVM) were built using the Python machine learning toolbox scikit-learn 0.19.1 (<http://scikit-learn.org/>), and the grid search method was used for parameter optimization. The eightfold cross-validation was applied to evaluate the performance of TML. The eightfold dataset was set to be the same as in DL training and testing for a fair comparison of G1/2 and G3 prediction accuracy between TML and DL.

Prediction of the Pathological Grading of pNENs by Radiologists

In order to compare the performance of DL, TML and radiologists in predicting the pathological grading of pNENs, the cases used in the TML models were evaluated by the radiologists.

Two abdominal radiologists – S. C and S. J – with 5 and 15 years of experience, respectively, who were blinded to both the presence and grade of pNENs, measured and recorded various CT features in accordance to the study by Kim et al. [10]. Any conclusive discrepancies between the radiologists were resolved by consensus after a joint re-assessment of the images. A *t* test was used for comparison of continuous variables and a χ^2 test or Fisher's exact test was used for comparison of categorical variables. The determination of the optimal cutoff values was same with that of Kim et al. [10].

Differences between DL, TML, and radiologists with regard to the accuracy of differentiating G3 from G1/2 tumors were also analyzed with a χ^2 test.

Table 3. Predicting the pathological grading of pNENs by using DL and TML models (91 patients from Hospital I)

	Accuracy, %	Sensitivity, %	Specificity, %	AUC	<i>p</i> value
Deep learning	72.5	81.3	70.7	0.80 (0.70–0.88)	–
Logistic regression	72.5	75.0	72.0	0.72 (0.62–0.81)	0.36
Random forest	70.3	75.0	69.3	0.72 (0.62–0.81)	0.24
SVM	71.4	87.5	68.0	0.80 (0.71–0.88)	0.93

The 95% CI is shown in the brackets. *p* value indicates the significance level of comparison of AUC with that of the DL model. Note that with deep learning method the AUC was 0.8 but not 0.81 as in Table 2, because 2 patients were excluded for this comparison between TML and DL due to the unclear boundary of the lesions. pNENs, pancreatic neuroendocrine neoplasms; AUC, area under the curve; TML, traditional machine learning; DL, deep learning; SVM, support vector machine.

Validation with DL on the Independent Testing Set

To further evaluate the robustness of the DL method, the CECT images of the 19 patients from Hospital II were used as an independent testing set. With the images of the optimal phase, the best model from the eightfold cross-validation, given in Section 4.3, was selected for this independent testing. The testing result was assessed using the ROC curve analysis.

Survival Analysis

Survival analysis was performed to explore the potential of the tumor pathological grade classifier in survival prediction. Patients from the 2 hospitals were divided into the G1/2 and G3 group according to the prediction results of the DL method. The Kaplan-Meier (KM) method was used for the survival analysis of G1/2 and G3 pNEN groups. Progression free survival (PFS) was defined as the time from the date of surgery to the date when progressive or recurrent disease was detected on imaging or to the last CT scan for those who had no tumor progression according to RECIST version 1.1 [22]. Overall survival (OS) was defined as the time from the date of surgery to the time until death. Alive patients were censored at the time of last follow-up.

Results

Performance of DL in Predicting the Pathological Grading of pNENs with Different Phases

Table 2 lists the accuracy, sensitivity, specificity, and AUC in predicting the pathological grading of the pNENs with DL method. The accuracy and AUC of the arterial model were higher than those of the venous and arterial/venous models. The AUC of the arterial model (0.81) was significantly higher than those of the venous model (AUC 0.57, *p* = 0.03) and arterial/venous model (AUC 0.70, *p* value = 0.03). Figure 4 shows the ROC curves of all the models.

Comparison between DL and TML

As the accuracy and AUC of the arterial model was higher than that of the venous and arterial/venous models

in predicting the pathological grading of the pNENs, the CECT images of the arterial phase were selected to build the TML models. Among the 93 patients with pNENs from Hospital I, the CECT images of the arterial phase showed unclear boundaries in 2 patients and hence the ground truth could not be placed. Thus, 91 cases (G1/2: *n* = 75, G3: *n* = 16) were included to build the TML models.

Table 3 lists the results. The DL model showed a higher AUC compared with the TML methods. Further pairwise comparisons revealed that the AUC difference between the DL and TML models was statistically insignificant (DL vs. LR, *p* = 0.36; DL vs. RF, *p* = 0.24; DL vs. SVM, *p* = 0.93).

Prediction of the Pathological Grading of pNENs by Radiologists

The 91 cases (G1/2: *n* = 75, G3: *n* = 16) used in the TML models were evaluated by the radiologists. The CT findings were summarized in Table 4. The results showed that in terms of the tumor size, poorly defined margin, vascular invasion, distant metastasis, arterial enhancement ratio (A ratio), portal venous enhancement ratio (P ratio), as well as non-expression of somatostatin receptors, there was a statistically significant difference between G1/2 and G3 pNENs.

The sensitivity, specificity, and OR of each cutoff value of the significant continuous variables (tumor size, A ratio and P ratio) for differentiating G3 from G1/2 tumors are summarized in Table 5.

The sensitivity and specificity of each significant CT criterion for differentiating G3 from G1/2 tumors are summarized in Table 6, which showed the highest OR for *p* ratio < 0.9. When at least 2 of these above 6 parameters were combined, the sensitivity for diagnosing G3 tumors was 100.0% (16/16), but the specificity was only 50.7% (38/75). When at least 3 of these 6 parameters were com-

Table 4. Clinical and pathological findings in G1/2 and G3 pNENs (91 patients from Hospital I)

	G1/G2 (<i>n</i> = 75)	G3 (<i>n</i> = 16)	χ^2/t value	<i>p</i> value
Age, years	46.81±13.48	49.13±17.21	−0.592	0.555
Size, mm	31.73±16.48	48.44±28.83	−3.171	0.002
Gender				
Male	36	8	0.021	0.884
Female	39	8		
Presentation				
Abdominal pain	27	8	2.987	0.810
Hypoglycemia	20	2		
No discomfort	12	2		
Weight loss	3	0		
Diarrhea	3	1		
Jaundice	6	2		
Poor appetite	4	1		
Location				
Head	28	8	0.909	0.635
Body	25	4		
Tail	22	4		
Pattern				
Solid	61	12	5.241	0.155
Solid and cystic	12	3		
Complex cystic	2	0		
Cystic	0	1		
Poorly defined margin	15	10	11.95	0.001
Calcification	12	1	–	0.451
Pancreatic duct dilation	23	9	3.786	0.052
Bile duct dilation	10	4	–	0.260
Pancreatic atrophy	21	6	0.570	0.450
Vascular invasion	25	10	1.585	0.029
Local lymphadenopathy	17	5	0.530	0.467
Distant metastasis	16	11	14.208	<0.001
Functional tumor	23	3	–	0.543
Non-SSTR expression	26	10	4.273	0.039
Arterial enhancement ratio, %	109.54±35.06	76.23±28.49	3.553	0.001
Venous enhancement ratio, %	110.45±25.35	83.42±25.62	3.865	<0.001

pNENs, pancreatic neuroendocrine neoplasms; SSTR, staining of somatostatin receptor.

bined, the sensitivity and specificity for diagnosing G3 tumors were 81.3% (13/16) and 72.0% (54/75) respectively. Table 6 lists the difference between the AUC of DL, TML and radiologists in differentiating G3 from G1/2 tumors, which were statistically insignificant.

Performance of DL on the Independent Testing Set

With the CECT images of the arterial phase, the best model from the eightfold validation was selected to evaluate the pathological grading of the 19 pNEN patients from Hospital II in the independent test set. The results showed the accuracy, sensitivity, specificity, and AUC of the DL algorithm were 82.1, 88.3, 84.6, and 0.82% respectively (Fig. 5).

Survival Analysis

KM analysis (Fig. 6) showed significant difference between the DL-predicted G1/2 group and the G3 pNEN group in both PFS ($p = 0.001$) and OS ($p < 0.001$), which suggested the prognostic value of the DL model.

Discussion

In this study, we applied a CNN-based DL method to predict the pathological grading of pNENs from CECT images. The established DL model showed a robust performance on an independent testing set. When predicting the pathological grading of pNENs by radiologists,

Table 5. Cutoff values of size and enhancement ratio for differentiating G3 from G1/2 pNENs (91 patients from Hospital I)

	Accuracy, %	Sensitivity, %	Specificity, %	AUC	OR	95% CI
Size, cm						
>2	42.9	93.8	32.0	0.629	7.059	0.880–56.591
>3	51.6	68.8	48.0	0.584	2.031	0.643–6.414
>4	64.8	50.0	68.0	0.590	2.125	0.712–6.343
Enhancement ratio						
A ratio						
<0.9	68.1	75.0	66.7	0.708	6.000	1.755–20.517
<1.1	58.2	87.5	52.0	0.698	7.583	1.611–35.704
p ratio						
<0.9	79.1	81.3	78.7	0.800	15.979	4.054–62.982
<1.1	58.2	93.8	50.7	0.722	15.405	1.936–122.604

A ratio, Arterial enhancement ratio; p ratio, portal venous enhancement ratio; pNENs, pancreatic neuroendocrine neoplasms; AUC, area under the curve.

Table 6. Significant clinico-pathological findings of for differentiating G3 from G1/2 pNENs (91 patients from Hospital I)

CT findings	Accuracy, %	Sensitivity, %	Specificity, %	AUC	OR	95% CI	p value (vs. DL)	p value (vs. SVM)	p value (vs. RF)	p value (vs. LR)
Poorly defined margin	76.9	62.5	80.0	0.603	2.455	0.929–6.493	0.495	0.397	0.313	0.495
Vascular invasion	65.9	62.5	66.7	0.646	3.333	1.087–10.218	0.335	0.424	0.525	0.335
Distant metastasis	76.9	68.8	78.7	0.737	8.112	2.461–26.741	0.495	0.397	0.313	0.495
Non-SSTR expression	64.8	62.5	65.3	0.639	0.318	0.104–0.974	0.156	0.210	0.276	0.156
Tumor size >4 cm	64.8	50.0	68.0	0.590	2.125	0.712–6.343	0.263	0.340	0.428	0.263
P ratio <0.9	79.1	81.3	78.7	0.800	15.979	4.054–62.982	0.192	0.079	0.057	0.057
Findings, n										
≥2	59.3	100	50.7	0.753	6.986E8	0.000	0.061	0.087	0.121	0.061
≥3	73.6	81.3	72.0	0.766	11.143	2.881–43.103	0.867	0.740	0.620	0.867
≥4	80.2	50.0	86.2	0.683	6.500	1.988–21.254	0.222	0.116	0.122	0.222

pNENs, pancreatic neuroendocrine neoplasms; CT, computed tomography; AUC, area under the curve DL, deep learning; SVM, support vector machine; RF, random forest; LR, logistic regression; p ratio, portal venous enhancement ratio; SSTR, staining of somatostatin receptor.

our results revealed that the highest OR was achieved for p ratio <0.9, which is different with that of Kim et al. [10] who showed that the highest OR was achieved for the p ratio <1.1. The results of the other significant criterion in our study for diagnosing G3 pNENs were also different with that of others [10, 12]. This suggested that when the manually measured CT features were used to differentiate G3 from G1/2 pNENs, the results from different institutions may be different. Therefore, the manually measured criterion from one institution cannot be used as reference in another institute. However, such limitation for manually measured criterion was not observed with the DL method. As shown in Figure 5, the AUC of the DL method was 0.82 in differentiating G3 from G1/2

tumors of independent testing set from Hospital II, which was stable compared to the performance with the data from Hospital I. Although the scanning protocol of independent testing set from Hospital II is different from the dataset from Hospital I, a robust accuracy was still demonstrated on the DL method. There may be 2 explanations for this finding. First, CT images are known to be linear to the tissue density and the difference between CT values of difference tissues may not change significantly between scanners/protocols; thus, the image features remain stable to some extent. Second, the image features extracted by the DL method are objective, complex, and abstract; they are less affected by the scanners/protocols.

Our study suggests that the AUC of the arterial phase was significantly higher than that of the venous and arterial/venous phases. This may be because the difference in the CT image features between G1/2 and G3 was greater during the arterial phase than that in the venous and arterial/venous phases. A previous report showed that the enhancement degree of G1/2 pNEN was significantly higher than that of G3 pNEN during the arterial phase [9]. However, there was no significant difference between G1/2 and G3 in terms of the enhancement degree of the venous phase [9]. When the venous and arterial phases were used in combination, the accuracy of the arterial/venous phase was not as good as the one obtained using the arterial phase alone. Theoretically, the arterial/venous dataset contains more information than the arterial or portal venous datasets. Therefore, the features extracted from the arterial/venous images by DL method may be more complex than the other 2 datasets. During training and validation processes with the 3 datasets (arterial, venous, and arterial/venous), the same eightfold cross-validation and network architecture were used. The arterial model has shown the best performance compared to the other 2 models, which proved that the feasibility of arterial images in predicting the pathological grading of pNENs with the DL method. In comparison, the venous model has shown the worst performance. This may be due to the image features extracted during the venous phase using DL are not distinguishable in differentiating G1/2 from G3 pNENs, which it would further increase the invalid information in the arterial/venous model, resulting in a worse performance of arterial/venous model compare to that of arterial model.

TML was also used to predict the pathological grading of pNENs. The results revealed that the DL exhibited higher performance to some extent, although the difference in AUC was insignificant. Compared to TML, DL methods may have some advantages. First, TML requires manual contouring of the tumor boundary on the images, which can be tedious and time-consuming. The DL-based models do not require drawing complex tumor boundaries, and the important features can be automatically learned [23–25]. Although it takes a long time to train a CNN model, the classification can be performed quickly once the DL model is established [25, 26]. In contrast, TML takes much less time for training, but it requires accurate tumor contouring which is time consuming for radiologists. Second, the CNN-based DL algorithm extracts the image features from the whole volume, while radiomics methods with TML only focus on the image patterns within the tumor lesions. Thus, the patients who

Fig. 5. ROC curves of the best DL model applied to the independent testing set from Hospital II. AUC, area under the curve.

have lesions with unclear boundaries cannot be analyzed with TML methods, whereas all the cases can be evaluated with the DL method. Third, compared to the image features extracted by the radiomics toolbox and then were selected by using 2-sample *t* test or Mann-Whitney U test, the features acquired by the DL method would be more abstract and stable without hypothesis, since they are the nonlinear combinations of features that are achieved after several convolution and nonlinear activation operations. It is reasonable to assume that the DL is more generalizable and accurate than TML.

It should be noted that both DL and TML have inherent technical limitations. With radiomics, since thousands of features are tested at the same time, multiple testing correction was not performed in most published studies; thus, some features may be false positives, and overfitting often occurs especially when the number of features is far greater than the number of patients in the dataset [27, 28]. Also, radiomics features are handcrafted and many of these features are similar, making the significant multicollinearity among these features. With DL methods, millions of parameters need to be tuned, and this requires a large sample [28]. In this study, we applied data augmentation to our training dataset for DL model training, and achieved relatively high and robust performance in both cross-validation and independent testing set. Although the features extracted by convolutional operations in DL method would be more stable, and may be less affected by the image difference caused by different

Fig. 6. Survival analysis using the real grades and DL predicted pNENs grades. The KM analysis with log-rank test shows significant difference between the predicted G1/2 group and G3 group in the PFS (**a**, $p = 0.001$) and OS (**b**, $p < 0.001$). OS, overall survival; PFS, progression free survival.

scanning parameters, the architecture of CNN is complex and should be specifically designed for different tasks [29]. In this study, we have adjusted the architecture of CNN (Res50 in our study) and optimized the network architecture to make it more suitable for our dataset. These included the modification of the Res50 network into 3D architecture, and setting the filter size as the same as the stride in order to avoid the image information loss. We have also tried to provide as much information as possible to the network, such as using the zero padding.

At present, it is difficult to precisely correlate the pathologic features, such as mitotic and Ki67 index (the standard reference for pathological grading of pNENs), with the CT image features by using DL, TML, or imaging features observed by the radiologists. In the current study, the researchers mainly used the imaging features extracted by various methods (such as DL, radiomics or imaging features observed by the radiologists), to establish a model to predict the pathological grading of pNENs [10–12]. No previous studies have been done to analyze the precise correlation between the pathological feature and the CT images. Jardim-Perassi et al. [30] have successfully demonstrated that multiple parametric MRI can identify distinguished hypoxic, viable, and non-viable tumor habitats in breast cancer mouse models. They have

also reported that the dynamic contrast-enhanced MRI demonstrated great potentials in identifying the hypoxic fractions in breast tumors. Similarly, we may assume that there are similar correlations between the pathological features such as mitotic count/Ki67 index and the CT/MR features, which is one of the aims for future research studies.

The KM survival analysis showed that the survival curves predicted by the DL model agreed well with the real survival curves derived from the patients' follow-up data. There was a significant difference in the DL model-predicted survival curves between G1/2 and G3 patients in PFS and OS, and this implied the capability of the DL model as a promising prognostic biomarker.

There may be some limitations to this study. First, our DL algorithm was based on a manual localization (semi-automatic) of the tumor lesions, and the accuracy would decline if the tumor lesions were incorrectly located. Nevertheless, in the diagnosis and treatment of pNENs, tumor lesions are routinely located, with rare incorrect localizations. Besides, studies have shown that the automatic segmentation of pancreas by DL has achieved encouraging performance, which may indicate that the automated segmentation of pNEN lesions is feasible and this may further automate the entire process of tumor

grading prediction. Second, the accuracy and AUC of the optimal CT images were 73.1 and 0.81%, respectively, which should be improved to meet the requirements of clinical practice. Except for the insufficient soft tissue contrast of the CT images, the network structure or the network parameters can be further fine-tuned to improve the prediction performance. Furthermore, the uneven distribution and limited size of the samples (93 pNEN patients from our institution: G1/2, $n = 76$; G3, $n = 17$) may reduce the prediction efficiency of the model. A larger sample was required for further verification of our proposed method.

In summary, by applying the proposed CNN-based DL model to CECT images preoperatively, prediction of pathological grading of pNENs is possible with a relatively high accuracy. The current model showed a robust performance on an independent testing set. CECT images were obtained from routine clinical practice, indicating that the DL classifiers can be efficiently implemented in current clinical practice, thus contributing to precision medicine applications. Further improvement is warranted to achieve a more automatic process and more accurate grading of pNENs when compared to the existing patterns.

Acknowledgments

Dr. Ling Xue and Yuan Lin (expert in pathology, Sun Yat-Sen University, China) provided pathological analysis for the manuscript.

References

- Halfdanarson TR, Rabe KG, Rubin J, Petersen GM. Pancreatic neuroendocrine tumors (PNETs): incidence, prognosis and recent trend toward improved survival. *Ann Oncol*. 2008 Oct;19(10):1727–33.
- Kloppel G. Classification and pathology of gastroenteropancreatic neuroendocrine neoplasms. *Endocr Relat Cancer*. 2011 Oct;18 Suppl 1:S1–16.
- Klimstra DS, Modlin IR, Coppola D, Lloyd RV, Suster S. The pathologic classification of neuroendocrine tumors: a review of nomenclature, grading, and staging systems. *Pancreas*. 2010 Aug;39(6):707–12.
- Lloyd RV, Osamura RY, Klöppel G, Rosai J. *WHO Classification of Tumours of Endocrine Organs*. 4th ed. Lyon, France: IARC Press; 2017.
- Basturk O, Yang Z, Tang LH, Hruban RH, Adsay V, McCall CM, et al. The high-grade (WHO G3) pancreatic neuroendocrine tumor category is morphologically and biologically heterogeneous and includes both well differentiated and poorly differentiated neoplasms. *Am J Surg Pathol*. 2015 May;39(5):683–90.
- Plöckinger U, Rindi G, Arnold R, Eriksson B, Krenning EP, de Herder WW, et al; European Neuroendocrine Tumour Society. Guidelines for the diagnosis and treatment of neuroendocrine gastrointestinal tumours. A consensus statement on behalf of the European Neuroendocrine Tumour Society (ENETS). *Neuroendocrinology*. 2004;80(6):394–424.
- Smith JK, Ng SC, Hill JS, Simons JP, Arous EJ, Shah SA, et al. Complications after pancreatic resection for neuroendocrine tumors: a national study. *J Surg Res*. 2010 Sep;163(1):63–8.
- Rossi S, Viera FT, Ghittoni G, Cobianchi L, Rosa LL, Siciliani L, et al. Radiofrequency ablation of pancreatic neuroendocrine tumors: a pilot study of feasibility, efficacy, and safety. *Pancreas*. 2014 Aug;43(6):938–45.
- Luo Y, Dong Z, Chen J, Chan T, Lin Y, Chen M, et al. Pancreatic neuroendocrine tumours: correlation between MSCCT features and pathological classification. *Eur Radiol*. 2014 Nov;24(11):2945–52.
- Kim DW, Kim HJ, Kim KW, Byun JH, Song KB, Kim JH, et al. Neuroendocrine neoplasms of the pancreas at dynamic enhanced CT: comparison between grade 3 neuroendocrine carcinoma and grade 1/2 neuroendocrine tumour. *Eur Radiol*. 2015 May;25(5):1375–83.
- Liang W, Yang P, Huang R, Xu L, Wang J, Liu W, et al. A combined nomogram model to preoperatively predict histologic grade in pancreatic neuroendocrine tumors. *Clin Cancer Res*. 2019 Jan;25(2):584–94.
- Canellas R, Burk KS, Parakh A, Sahani DV. Prediction of pancreatic neuroendocrine tumor grade based on CT features and texture analysis. *AJR Am J Roentgenol*. 2018 Feb;210(2):341–6.

Statement of Ethics

All subjects enrolled in this study have provided written informed consent for the research study protocol. All methods were carried out in accordance with the approved guidelines. The ethics approval was provided by the First Affiliated Hospital, Sun Yat-Sen University, China.

Disclosure Statement

The authors declare that they have no conflicts of interest to disclose.

Funding Sources

This work was funded by National Natural Science Foundation of China (81971684, 81771908, 81571750, 81770654, 81801761, 81471735), National Key Research and Development Program of China (2017YFC0113402), Guangzhou Science and Technology Foundation (201804010078), Shenzhen Municipal Scheme for Basic Research (JCYJ20160428164548896).

Author Contributions

S.-T.F., B.H., Y.L., and J.C. conceived the study. Y.L. and X.C. collected the data and wrote the main manuscript. J.C., X.C., and H.X. performed the data analysis. J.S., C.S., M.C., and Z.-P.L. completed the literature search, edited the manuscript, and all authors have read and approved the manuscript. Y.L., X.C., and J.C. contributed equally to this work in study design, data collection, interpretation, and manuscript drafting.

- 13 Wu G, Chen Y, Wang Y, Yu J, Lv X, Ju X, et al. Sparse Representation-Based Radiomics for the Diagnosis of Brain Tumors. *IEEE Trans Med Imaging*. 2018 Apr;37(4):893–905.
- 14 Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017 Aug;284(2):574–82.
- 15 Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology*. 2018 Mar;286(3):887–96.
- 16 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al.; the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017 Dec;318(22):2199–210.
- 17 Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, et al. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin Cancer Res*. 2018 Mar;24(5):1073–81.
- 18 Esteve A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017 Feb;542(7639):115–8.
- 19 Tamburrino D, Spoletini G, Partelli S, Muffatti F, Adamenko O, Crippa S, et al. Surgical management of neuroendocrine tumors. *Best Pract Res Clin Endocrinol Metab*. 2016 Jan;30(1):93–102.
- 20 Crippa S, Zerbi A, Boninsegna L, Capitanio V, Partelli S, Balzano G, et al. Surgical management of insulinomas: short- and long-term outcomes after enucleations and pancreatic resections. *Arch Surg*. 2012 Mar;147(3):261–6.
- 21 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; p. 770–8.
- 22 Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009 Jan;45(2):228–47.
- 23 Li R, Zhang W, Suk HI, Wang L, Li J, Shen D, et al. Deep learning based imaging data completion for improved brain disease diagnosis. *International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer*, 2014; p. 305–12.
- 24 Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. *AMIA Annual Symposium Proceedings: American Medical Informatics Association*, 2015; p. 1899.
- 25 Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. *Jpn J Radiol*. 2018 Apr;36(4):257–72.
- 26 Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoteric acid-enhanced hepatobiliary phase MR images. *Radiology*. 2018 Apr;287(1):146–55.
- 27 Wu G, Chen Y, Wang Y, Yu J, Lv X, Ju X, et al. Sparse Representation-Based Radiomics for the Diagnosis of Brain Tumors. *IEEE Trans Med Imaging*. 2018 Apr;37(4):893–905.
- 28 Zhang Y, Lobo-Mueller EM, Karanickolas P, Gallinger S, Haider MA, Khalvati F. Improving Prognostic Value of CT Deep Radiomic Features in Pancreatic Ductal Adenocarcinoma Using Transfer Learning. *arXiv*. 2019:1905.09888.
- 29 Hussein S, Kandel P, Bolan CW, Wallace MB, Bagci U. Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. *IEEE Trans Med Imaging*. 2019 Aug;38(8):1777–87.
- 30 Jardim-Perassi BV, Huang S, Dominguez-Viqueira W, Poleszczuk J, Budzevich MM, Abdalah MA, et al. Multiparametric MRI and Coregistered Histology Identify Tumor Habits in Breast Cancer Mouse Models. *Cancer Res*. 2019 Aug;79(15):3952–64.