Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation

Kao-Lang Liu*, Tinghui Wu*, Po-Ting Chen, Yuhsiang M Tsai, Holger Roth, Ming-Shiang Wu, Wei-Chih Liao, Weichung Wang

Summary

Background The diagnostic performance of CT for pancreatic cancer is interpreter-dependent, and approximately 40% of tumours smaller than 2 cm evade detection. Convolutional neural networks (CNNs) have shown promise in image analysis, but the networks' potential for pancreatic cancer detection and diagnosis is unclear. We aimed to investigate whether CNN could distinguish individuals with and without pancreatic cancer on CT, compared with radiologist interpretation.

Methods In this retrospective, diagnostic study, contrast-enhanced CT images of 370 patients with pancreatic cancer and 320 controls from a Taiwanese centre were manually labelled and randomly divided for training and validation (295 patients with pancreatic cancer and 256 controls) and testing (75 patients with pancreatic cancer and 64 controls; local test set 1). Images were preprocessed into patches, and a CNN was trained to classify patches as cancerous or non-cancerous. Individuals were classified as with or without pancreatic cancer on the basis of the proportion of patches diagnosed as cancerous by the CNN, using a cutoff determined using the training and validation set. The CNN was further tested with another local test set (101 patients with pancreatic cancers and 88 controls; local test set 2) and a US dataset (281 pancreatic cancers and 82 controls). Radiologist reports of pancreatic cancer images in the local test sets were retrieved for comparison.

Findings Between Jan 1, 2006, and Dec 31, 2018, we obtained CT images. In local test set 1, CNN-based analysis had a sensitivity of 0·973, specificity of 1·000, and accuracy of 0·986 (area under the curve [AUC] 0·997 [95% CI 0·992–1·000]). In local test set 2, CNN-based analysis had a sensitivity of 0·990, specificity of 0·989, and accuracy of 0·976 (AUC 0·999 [0·998–1·000]). In the US test set, CNN-based analysis had a sensitivity of 0·790, specificity of 0·992, and accuracy of 0·992 (AUC 0·999 [0·998–1·000]). In local test set 2, CNN-based analysis achieved higher sensitivity than radiologists did (0·983 vs 0·929, difference 0·054 [95% CI 0·011–0·098]; p=0·014) in the two local test sets combined. CNN missed three (1·7%) of 176 pancreatic cancers (1·1–1·2 cm). Radiologists missed 12 (7%) of 168 pancreatic cancers (1·0–3·3 cm), of which 11 (92%) were correctly classified using CNN. The sensitivity of CNN for tumours smaller than 2 cm was 92·1% in the local test sets and 63·1% in the US test set.

Interpretation CNN could accurately distinguish pancreatic cancer on CT, with acceptable generalisability to images of patients from various races and ethnicities. CNN could supplement radiologist interpretation.

Funding Taiwan Ministry of Science and Technology

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction

Pancreatic cancer is projected to become the second leading cause of cancer deaths in the USA by 2030,1 with dismal survival once the tumour size exceeds 2 cm.2 CT is the major imaging modality used for detection and assessment of pancreatic cancer,3 but the method’s diagnostic performance depends on radiologists’ experience. Furthermore, approximately 40% of tumours that are smaller than 2 cm evade detection by CT,4 underscoring an urgent need for novel methods to supplement radiologist interpretation in improving the sensitivity for the detection of pancreatic cancer.

Deep learning with convolutional neural networks (CNNs) has shown great promise in medical image analysis.5 The construction of neural networks is based on a stack of neurons composed of activation functions and parameters to extract and integrate features from the images and establish a model that captures the complex relationship between images and diagnoses. CNN has been reported to achieve a high accuracy in the imaging diagnosis of various conditions including skin cancer,6 diabetic retinopathy,7 and liver masses.8 However, the potential usefulness of CNN for the detection and diagnosis of pancreatic cancer has not been widely investigated. Most pancreatic cancers present with irregular contours and ill-defined margins on CT and thus are often obscure at an early stage, posing substantial challenges even for the most experienced radiologists.9,10
How well a CNN can capture the elusive CT features of pancreatic cancer and be used to detect or diagnose pancreatic cancer is unclear.

Ascertaining the potential usefulness of CNN in detecting and diagnosing pancreatic cancer has important implications. We hypothesised that with high-quality training data and image preprocessing that is tailored to the characteristics of pancreatic cancer, CNN could effectively be used to distinguish pancreatic cancer from non-cancerous pancreas. To provide a proof of concept, we aimed to train a CNN-based classification model for differentiating between pancreatic cancer and non-cancerous pancreas on CT and assess the method’s performance using local and external test datasets with comparison with radiologists’ reports.

Methods
Study design
We did a retrospective, diagnostic study, using CT images from the imaging archive of National Taiwan University Hospital (NTUH), the Medical Segmentation Decathlon dataset, and The Cancer Imaging Archive (TCIA) dataset.

This study was approved by the Research Ethics Committee of NTUH, a tertiary referral centre with many patients with pancreatic cancer. Informed consent from individual patients was waived because of the retrospective design of the study. Patients or the public were not involved in the design, or conduct, or reporting, or dissemination of our research. Detailed methods of image processing and analyses are provided in the appendix (p 3).

Evidence before this study
We searched PubMed from inception to Jan 4, 2020, for original research articles, using the search terms “pancreatic cancer convolutional neural network” and “pancreatic cancer deep learning”, without language restrictions. All the identified references were reviewed.

Convolutional neural network (CNN) has been reported for segmentation of pancreas, risk stratification of intraductal papillary mucinous neoplasms, prediction of grading of pancreatic neuroendocrine neoplasms, targeting pancreatic tumour for radiotherapy, and classification of pancreatic cysts. A case study reported the initial experience of using deep learning algorithms for the classification of patients with pancreatic cancer and healthy controls, without formal testing of the algorithm in independent datasets. The study analysed only a subset of available patients without describing the inclusion and exclusion criteria, and no information was provided regarding the reference standard used and the interval between the index test and reference standard. Therefore, that study is at risk of bias according to the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Added value of this study
We trained a CNN using contrast enhanced-CT images of Asian patients to distinguish pancreatic cancer from healthy pancreases. CNN achieved excellent accuracy and improved sensitivity compared with radiologist interpretation in independent Asian test sets, with acceptable performance in a North American test set obtained from patients of various races and ethnicities using diverse scanners and settings. These results provide the first solid proof of concept that CNN can capture the elusive CT features of pancreatic cancer to assist and supplement radiologists in the detection and diagnosis of pancreatic cancer.

Implications of all the available evidence
CNN can accurately differentiate pancreatic cancer from non-cancerous pancreas, and with improvements might accommodate variations in patient race and ethnicity and imaging parameters that are inevitable in real-world clinical practice. CNN holds promise for developing computer-aided detection and diagnosis tools for pancreatic cancer to supplement radiologist interpretation.

CT image datasets and radiologist reports
Patients with histologically confirmed or cytologically confirmed pancreatic adenocarcinoma were identified from the NTUH Cancer Registry, and the CT images of those patients obtained before the date of pancreatic cancer diagnosis were extracted from the imaging archive of NTUH for review. Contrast-enhanced portal venous CT images of 370 pancreatic cancer patients diagnosed between Jan 1, 2013, and Dec 31, 2018, and 320 patients with normal pancreas during the same period were randomly selected for training and testing of the deep learning model. For patients who had undergone more than one CT examination, only the examination that immediately preceded the date of diagnosis was selected. The control participants were selected on the basis of the statement of a negative or unremarkable pancreas in the original radiologist report and verified during manual labelling of the pancreas by one of the two radiologists participating in this study. The dataset was then randomly split into the training and validation set (220 patients with pancreatic cancer and 192 controls in the training set, and 75 patients with pancreatic cancer and 64 controls in the validation set) and a test set (75 patients with pancreatic cancer and 64 controls [local test set 1]). CT images of another 101 patients with pancreatic cancer and 88 controls were randomly selected from the NTUH imaging archive between Jan 1, 2006, and Dec 31, 2012, and used as the second test set [local test set 2; figure 1]. Among the 471 pancreatic cancer patients of NTUH who were included in the study, pancreatic cancer was diagnosed by histology in 355 (75·4%) patients and by cytology in
the others. The median interval between CT examination and histological or cytological diagnosis of pancreatic cancer was 16 days (IQR 8–30 days). The formal radiologist reports of the CT images of pancreatic cancer patients in the two local tests were retrieved from the NTUH electronic health record. After removal of the identity of the patient and interpreting radiologist, the radiologist reports were jointly reviewed by two radiologists (P-TC [5 years’ experience] and K-LL [20 years’ experience]) by one of two experienced abdominal radiologists (P-TC [5 years’ experience] and K-LL) without reference with the CT images. A radiologist report was considered to have correctly classified the patient as having pancreatic cancer if a description of a definite or suspicious pancreatic tumour was found in the report; otherwise the cancer was considered to be missed by the interpreting radiologist.

CT examinations were done using six scanners (Brilliance iCT 256, Philips Healthcare [Best, Netherlands]; Sensation 64 and SOMATOM Definition AS+, Siemens Healthcare [Forchheim, Germany]; Aquilion one, Toshiba [Tochigi, Japan]; Revolution CT and LightSpeed VCT, GE Medical system, [Milwaukee, WI, USA]) with 100, 120, 130 kV, or automatic mA control without extra noise reduction processes. The slice thickness was 0·7–1·5 mm, and window level as 75 HU. The images were manually labelled for further model training and validation and testing using an open source software (3D Slicer [version 4.8.1]) by one of two experienced abdominal radiologists (P-TC [5 years’ experience] and K-LL [20 years’ experience]). Because the pancreas bordered multiple organs and structures, and pancreatic cancers often had indistinct borders with the surrounding tissue, inter-observer differences might exist regarding the exact extent of the pancreas and the tumour. Therefore, the labelled pancreas and tumour on the images were jointly reviewed for consensus by the radiologists before further processing and analysis. The window width was set as 250 Hounsfield units (HU) and window level as 75 HU. The images were normalised to [0, 1] by linear interpolation, and the portions that were neither pancreas nor tumour were excluded from further analysis. The images were then cropped into square subregions (ie, patches) using the moving window method on the axial (x-y) plane, starting from the top-left corner. The window moved along the x-axis between the right and left boundaries, at which the window moved one step downward along the y-axis and then resumed moving along the x-axis toward the other border, until finally ending at the bottom-right corner. To increase the variation and size of training data, the moving distance was set as half of the patch dimension to generate overlapping patches. The patches that contained pancreatic cancer were labelled as cancerous, whereas patches that contained only non-cancerous pancreatic parenchyma were labelled as non-cancerous.

**Image preprocessing**

All axial CT images that contained the pancreas or pancreatic cancer from an individual participant were

---

**Figure 1: Local and external datasets for model training and testing**

| Local dataset—National Taiwan University Hospital image archive | External dataset |
|---------------------------------------------------------------|------------------|
| **Years 2013–18** | 281 patients with pancreatic cancer (4898 images from Memorial Sloan Kettering Center, New York, NY, USA) |
| 370 patients with pancreatic cancer (5657 images) | 82 patients with normal pancreas (1427 images) |
| 320 patients with normal pancreas (5900 images) | 101 patients with pancreatic cancer (1657 images) |
| **Years 2006–12** | **External test set** |
| 101 patients with pancreatic cancer (1566 images) | 281 patients with pancreatic cancer (4898 images) |
| 88 patients with normal pancreas (1657 images) | 82 patients with normal pancreas (1427 images) |
| **Local test set 1** | **Local test set 2** |
| 75 patients with pancreatic cancer (1181 images) | 101 patients with pancreatic cancer (1657 images) |
| 64 patients with normal pancreas (1177 images) | 88 patients with normal pancreas (1657 images) |
| **Validation set** | **Local dataset—National Taiwan University Hospital image archive** |
| 75 patients with pancreatic cancer and 64 with normal pancreas (2244 images) | **External dataset** |
| **Training set** | 101 patients with pancreatic cancer (1566 images) |
| 220 patients with pancreatic cancer and 192 with normal pancreas (5955 images) | 281 patients with pancreatic cancer (4898 images) |
| 75 patients with pancreatic cancer and 64 with normal pancreas (2244 images) | 82 patients with normal pancreas (1427 images) |

**Patch-based and patient-based analyses**

Using all patches of the training and validation set, a deep CNN was trained to identify the probability that a patch harbours pancreatic cancer and to classify the patch as
### Characteristics of patients with pancreatic cancer in local datasets

**Table 1:**

| Age, years | Training and validation set (n=295) | Test set 1 (n=75) | Test set 2 (n=101) | Combined (n=471) |
|------------|-----------------------------------|------------------|-------------------|------------------|
|            | 65.1 (12.2)                      | 63.9 (11.9)      | 65.9 (11.5)       | 64.8 (12.0)      |
| Sex        |                                   |                  |                   |                  |
| Male       | 156 (55%)                         | 40 (53%)         | 57 (56%)          | 260 (55%)        |
| Female     | 132 (45%)                         | 35 (47%)         | 44 (44%)          | 211 (45%)        |
| Tumour stage |                                   |                  |                   |                  |
| I          | 6 (2%)                            | 2 (3%)           | 9 (9%)            | 17 (4%)          |
| II         | 57 (33%)                          | 23 (31%)         | 47 (46%)          | 167 (35%)        |
| III        | 53 (18%)                          | 6 (8%)           | 15 (15%)          | 74 (16%)         |
| IV         | 129 (47%)                         | 44 (58%)         | 30 (30%)          | 212 (45%)        |
| Tumour size, cm | 3.3 (2.4±4.8)  | 3.4 (2.3±4.6)  | 2.5 (2.0±3.2) | 3.0 (2.2±4.6) |

Data are mean (SD), n (%), or median (IQR).

For comparison between groups, Fisher’s exact test was used for comparing categorical variables and Mann-Whitney U test for comparing continuous variables.

### Role of the funding source

The funder of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study, and W-CL and WW had final responsibility for the decision to submit for publication.

### Results

The clinical characteristics of the patients with pancreatic cancer in the local dataset are summarised in table 1. The dataset included 14780 CT images (7223 pancreatic cancers and 7557 controls). The training and validation set and two local test sets were similar with respect to demographic characteristics and tumour stage. Tumour size was treated as regions of interest (ROIs), which were processed into patches and subject to analysis by using the trained CNN. Patients were classified as with or without pancreatic cancer on the basis of the proportion of patches generated from the patient’s ROIs that were diagnosed as cancerous by the CNN model, using the cutoff that achieved the highest Youden index in the ROC curve that was constructed using the validation set.

### Model architectures and training

The CNN model was modified from Visual Geometry Group (VGG) network, a network widely used in image classification. First, the model contained three convolutional blocks. Each block consisted of two convolution layers followed by the rectified linear unit as the activation function, and then connected with a max-pooling layer before the next block. Thereafter, a flatten node was attached in the last convolution block, and three fully connected (dense) layers were attached at the end of the CNN model. Finally, weighted binary cross-entropy was used as the loss function to account for imbalance between the number of cancerous and non-cancerous patches. Details of the CNN, including the layer structures, kernel sizes, channels, and output sizes of the network are given in the appendix (p 5).

The process of model training and testing is summarised in the appendix (p 3). To optimise model performance, two callbacks based on validation loss were used during the training process. First, learning rate would be reduced to 10% if loss did not decrease for ten iterations. Second, early stopping callback would stop the training after the validation loss remained stable after 40 iterations to avoid overfitting. All codes were written in Python (version 3.6.8) using Keras (version 2.2.4) and Tensorflow (version 1.7.0) libraries.

### Model testing with local and external datasets

Model performance was assessed using local and external test sets to ascertain sensitivity, specificity, and accuracy with respect to patch-based and patient-based analyses. The external test set consisted of two open-source datasets from research institutions in the USA (figure 1). The Medical Segmentation Decathlon dataset from Memorial Sloan Kettering Cancer Center, New York, NY, USA included CT images of 281 patients with pancreatic cancer, and TCIA dataset included CT images of 82 individuals with normal pancreas from the US National Institutes of Health Clinical Center.

### Statistical analysis

For comparison between groups, Fisher’s exact test was used for comparing categorical variables and Mann-Whitney U test for comparing continuous variables.
The performance of the CNN in various datasets is summarised in table 3. The training and validation set yielded 244859 cancerous patches and 1216712 non-cancerous patches for model training and adjustment of learning rate. In the training and validation set, the trained CNN achieved an area under the curve (AUC) of 0·955 (95% CI 0·955–0·956) in differentiating between cancerous and non-cancerous patches (figure 2A), with a sensitivity of 0·913 (0·912 to 0·914), specificity of 0·856 (0·844 to 0·865), accuracy of 0·879, and balanced accuracy of 0·955 (0·955 to 0·956). In patient-based analysis, the sensitivity was 0·878 (0·876–0·881), specificity 0·850 (0·855 to 0·855), accuracy 0·856, and balanced accuracy 0·956 (0·955 to 0·956). The local test set 1 contained 1566 cancerous and 1657 non-cancerous CT images, yielding 57874 cancerous patches and 440982 non-cancerous patches. In patch-based analysis, the sensitivity was 0·875 (0·872–0·878), specificity 0·907 (0·906–0·907), accuracy 0·942 (0·941 to 0·943), and balanced accuracy 0·964 (0·963 to 0·965). In patient-based analysis, the sensitivity was 0·946 (0·945–0·947), specificity 0·989 (0·988–0·989), accuracy 0·989, and balanced accuracy 0·962 (0·961–0·962). For both local test sets combined, the patient-based sensitivity was 0·974 (0·973–0·975) and specificity 0·991 (0·990–0·991), with an accuracy of 0·986 (0·985–0·986) and balanced accuracy of 0·991 (0·990–0·991).

The local test set 1 consisted of 1181 cancerous and 1177 non-cancerous CT images, yielding 65075 cancerous patches and 306920 non-cancerous patches. In patch-based analysis, the sensitivity was 0·912 (0·910–0·915), specificity 0·858 (0·856–0·858), accuracy 0·867 (0·866–0·868), and balanced accuracy 0·885 (0·884–0·886). In patient-based analysis, the sensitivity was 0·973 (0·907–0·997), specificity 1·000 (0·944–1·000), accuracy 0·986 (0·949–0·998), and balanced accuracy 0·987 (0·968–1·000; figure 2D). The local test set 2 consisted of 1566 cancerous and 1657 non-cancerous CT images, yielding 57874 cancerous patches and 440982 non-cancerous patches. In patch-based analysis, the sensitivity was 0·875 (0·872–0·878), specificity 0·907 (0·906–0·907), accuracy 0·942 (0·941–0·943), and balanced accuracy 0·964 (0·963–0·965; figure 2E). In patient-based analysis, the sensitivity was 0·946 (0·945–0·947), specificity 0·989 (0·988–0·989), accuracy 0·989 (0·962–0·999), and balanced accuracy 0·989 (0·975–1·000; figure 2F). For both local test sets combined, the patient-based sensitivity was 0·983 (0·951–0·996) and specificity 0·993 (0·964–1·000), with an accuracy of 0·988 (0·969–0·997) and balanced accuracy 0·988 (0·977–1·000). Figure 3 shows the correspondence between model diagnoses and radiologist’s labelling with respect to tumour location.

In local test set 1, the sensitivity for detecting pancreatic cancer for CNN-based analysis was similar to that of radiologists (table 3). In local test set 2, in which participants had smaller tumours than did those in test set 1, CNN-based analysis achieved significantly higher sensitivity than that of radiologists (table 3). For both local test sets combined, the sensitivity of CNN-based analysis was also significantly higher than that of radiologists (0·983 [0·964–1·000] vs 0·929 [0·890–0·968]), with a difference of 0·054 (0·011–0·098; p=0·014). Three (1·7%) patients with pancreatic cancer...
were missed by CNN-based analysis, and 12 (7.1%, excluding eight patients without radiology reports) were missed by radiologists (appendix p 2). The three tumours missed by CNN were 1.1–1.2 cm in size (figure 4A, B), of which two were correctly classified by the radiologists. The 12 tumours missed by radiologists were 1.0–3.3 cm in size, of which 11 (92%) were correctly classified by CNN-based analysis (figure 4C, D). One tumour sized 1.2 cm was missed by both the CNN and the radiologist. The specificity of the radiologist reports was not assessed because the control participants were selected on the basis of the statement of a normal or unremarkable pancreas in the radiologist report.

The external test set consisted of 4898 cancerous and 1427 non-cancerous CT images, yielding 144 249 cancerous patches and 807 911 non-cancerous patches. The sensitivity, specificity, accuracy, and balanced accuracy of the trained CNN in differentiating between cancerous and non-cancerous patches and between participants with and without pancreatic cancer are given in figure 2G and H and table 3.

Because the major limitation of radiologist interpretation is inadequate sensitivity for small pancreatic cancers, we further compared the sensitivity of CNN-based analysis and radiologists in detecting pancreatic cancer stratified by cancer stages and categories of primary tumour size (table 4). For the two local test sets combined, the sensitivity of CNN-based analysis and radiologists was similar (0.921 [95% CI 0.786 to 0.983] vs 0.895 [0.752 to 0.971]), with a difference of 0.026 (−0.104 to 0.156; p=0.692) in tumours smaller than 2 cm (table 4). For tumours 2–4 cm in size, CNN-based analysis achieved higher sensitivity compared with that of radiologists (table 4; 1.000 [0.961 to 1.000] vs 0.908 [0.827 to 0.959], difference 0.092 [0.031 to 0.153]; p=0.003). For tumours larger than 4 cm, CNN-based analysis and radiologists achieved 100% sensitivity (table 4). In the external test set, the sensitivity of CNN-based analysis was 0.631 (0.502 to 0.747) for tumours smaller than 2 cm, 0.823 (0.760 to 0.875) for tumours 2–4 cm, and 0.933 (0.779 to 0.992) for tumours larger than 4 cm (p_{trend}<0.001). Information on cancer stage was not available in the external test set.

**Discussion**

This study provided a proof of concept that CNN-based analysis could accurately distinguish patients with and without pancreatic cancer on portal venous CT. The CNN-based analysis achieved an accuracy approaching 99% and missed fewer tumours compared with that of radiologists. Furthermore, the CNN that was trained using CT images of Asian patients with pancreatic cancer from a single institution achieved acceptable performance when tested by CT images obtained from patients of different races and ethnicities using diverse scanners and settings. These results support that CNN can robustly recognise the CT characteristics of pancreatic cancer and thus hold great promise for developing computer-aided detection and diagnosis tools for pancreatic cancer.

Although deep learning has been studied for the diagnosis of pancreatic cystic neoplasms, neuroendocrine tumours and segmentation of the pancreas, the technique’s potential usefulness in the diagnosis of pancreatic cancer, the most common pancreatic malignancy, has not been widely investigated. This study focused on examining whether CNN can tell the subtle differences between pancreatic cancer and normal pancreas that often evade the naked eye, because these differences are a potential window of opportunity to fulfil the utmost unmet need of improving the
sensitivity for detecting small pancreatic cancers. Major challenges exist in applying deep learning to the diagnosis of pancreatic cancer. First, 80% of pancreatic cancers have irregular contours and ill-defined margins on CT, leading to inconspicuous borders with surrounding tissue. Second, the fact that a tumour occupies only a portion of the pancreas, which is surrounded by multiple organs and structures, poses severe challenges in model fitting because the CNN might be distracted away from the tumour towards irrelevant regions, which comprise most of the image. Third, deep learning requires a large number of labelled images, but labelling of the pancreas and tumour is cumbersome and time consuming because of the organ’s anatomical complexity. To tackle these difficulties, we preprocessed CT images into patches and used techniques including moving window and flipping as a means of data augmentation. Our patch-based analytical approach also had the advantage of allowing the CNN to interrogate the tumour multiple times, rather than once only if the pancreas and tumour was analysed as a whole. Because the diagnosis of pancreatic cancer versus non-pancreatic cancer (ie, patient-based analysis) was based on analyses of multiple patches, the results might be less sensitive to variations in image quality that would affect the accuracy of patch-based analysis, and thus were more robust than results based on a single analysis of the pancreas and tumour as a whole. Therefore, the sensitivity and specificity of patient-based analyses could exceed those of patch-based analyses in all datasets in this study. We also excluded unrelated neighbouring organs and structures when training the CNN. This approach yielded significantly better performance compared with CNNs trained with neighbouring organs and structures included. The CNN-based analysis achieved almost 99% accuracy in multiple local test sets, with acceptable generalisability to the US test set, providing the first proof of concept that with proper image preprocessing, CNN can capture the complex imaging characteristics of pancreatic cancer. Notably, the CNN-based analysis was able to correctly identify the real site of the tumour, even in the presence of seemingly similar secondary findings such as a dilated pancreatic duct, which might seem similar to the tumour. The potential ability of CNNs in detecting and distinguishing various pancreatic diseases in addition to pancreatic cancer (eg, pancreatitis and other pancreatic tumours) should be explored in future research.

The neural network architecture used in this study was modified from the VGG network, a classic CNN architecture, which has been widely used for image detection and segmentation. The original VGG network was trained for object recognition and consisted of 13 convolution layers and three fully connected layers. The VGG network has been shown to achieve good performance on the ImageNet dataset and subsequently applied to medical image analysis in tasks such as classification of common diseases on chest x-ray, identification of metastatic breast cancer on histological evaluation of lymph node biopsies, and segmentation of blood vessels and optic disc in retinal imaging. However, the input size in the original VGG network (224×224 pixels) is substantially larger than the size of the patches in this study and thus might cause overfitting. Therefore, we reduced the number of convolutional layers from 13 to six to reduce the complexity of the model, and the finding that model performance in the local test sets was similar to that in the training and validation set suggested that no notable overfitting occurred.

The notable finding of the CNN achieving higher sensitivity compared with that of radiologists in the local test sets supports the potential usefulness of the CNN in...
Example cases of pancreatic cancers missed by CNN or radiologists

Figure 4: Examples of pancreatic cancers missed by CNN or radiologists

(A) and (B) show tumours missed by CNN but correctly classified by the radiologists; and (C) and (D) show tumours missed by the radiologists but correctly classified by CNN. (A) CT showed a pancreatic head nodule (arrowhead) in a 71-year-old woman with epigastric pain and obstructive jaundice, confirmed as adenocarcinoma by histology after surgical resection. (B) CT showed an irregular pancreatic head hypodense lesion (arrowhead) in an 88-year-old man with incidental findings of tail and a hypodense lesion at pancreatic body, which was missed by the radiologist and subsequently found to be pancreatic adenocarcinoma by histology after surgical resection. (C) CT of a 57-year-old man presenting with acute pancreatitis showed changes of acute pancreatitis at pancreatic tail and a hypodense lesion at pancreatic body, which was missed by the radiologist and subsequently found to be pancreatic adenocarcinoma by histology after surgical resection. (D) A 77-year-old man with incidental findings of dilated intrahepatic and extrahepatic bile ducts on CT. A pancreatic head hypodense lesion (arrowhead) was missed by the radiologist and subsequently proven to be pancreatic adenocarcinoma by histology after surgical resection.

CNN=convolutional neural network.

A trade-off generally exists between sensitivity and specificity, but an ideal computer-aided diagnostic tool for pancreatic cancer needs to have high sensitivity and specificity simultaneously. To that end, this study adopted a unique two-level analytical approach. If a single-level analysis was used, lowering the threshold to increase the sensitivity for detecting pancreatic cancer would assist radiologists and clinicians. Because the two local test sets and training and validation set shared no common patients, the good performance of the CNN could not be attributed to overfitting. Even in a tertiary medical centre, 7.1% of the pancreatic cancers were missed by radiologist interpretation. This finding highlights the common clinical scenario of pancreatic cancers being present on CT images but missed and only discovered in retrospect when the cancers have become clinically apparent and underscores the limitations in the diagnostic performance of CT for pancreatic cancer.

Patient survival dramatically decreases when tumour size exceeds 2 cm, but tumours smaller than 2 cm are often inconspicuous on CT, with approximately 40% of such tumours being missed. Disparities in radiologists’ expertise and experience and heavy workload might further adversely affect the accuracy of radiologist interpretation. Given the rising incidence of pancreatic cancer and dire consequences associated with missed or delayed diagnosis, solutions to these limitations are urgently needed. Our results suggest that with further improvements, a CNN-based computer-aided diagnostic or detection tool might reduce the miss rate in clinical practice and narrow disparities in resources and expertise.

### Table 4: Sensitivity in test sets stratified by tumour stage and size

| Stage | External test set (n=281) | Local test set 1 (n=75) | Local test set 2 (n=101) | Local test sets combined (n=176) | Difference | p value |
|-------|--------------------------|-------------------------|--------------------------|---------------------------------|------------|---------|
|       | By CNN                  | By radiologist          |                         |                                 |            |         |
| I     | 1000 (0.715 to 1.000); 1000 (0.664 to 1.000); 11/11 (100%) | 9/9 (100%)              |                         |                                 |            |         |
| II    | 0.957 (0.880 to 0.991); 6/7/0 (96%) | 46/47 (98%)             |                         |                                 |            |         |
| III   | 1000 (0.782 to 1.000); 15/15 (100%) | 1000 (0.939 to 1.000); 21/21 (100%) |                         |                                 |            |         |
| IV    | 1000 (0.928 to 1.000); 44/44 (100%) | 1000 (0.951 to 1.000); 74/74 (100%) |                         |                                 |            |         |
|       | By CNN                  | By radiologist          |                         |                                 |            |         |
| Size (cm) |                         |                         |                         |                                 |            |         |
| <2    | 0.014 (0.013 to 0.041); 0.009 (0.013 to 0.041); 70/71 (99%) | 79/87 (99%)             |                         |                                 |            |         |
| 2–4   | 0.996 (0.986 to 0.999); 26/27 (96%) | 0.921 (0.786 to 0.983); 35/38 (92%) |                         |                                 |            |         |
| >4    | 0.026 (0.104 to 0.156); 0.026 (0.104 to 0.156); 34/38 (99%) | 0.935 (0.786 to 0.971); 34/38 (99%) |                         |                                 |            |         |

Data are sensitivity (95% CI), or n/N (%). Only patients with pancreatic cancer were analysed. Sensitivity by stage in the external dataset could not be assessed because information on cancer stage was not available. CNN=convolutional neural network. *Radiologist report could not be identified in two patients. †Radiologist report could not be identified in 3 patients. ‡Radiologist report could not be identified in three patients. §Radiologist report could not be identified in 5 patients. ¶Radiologist report could not be identified in 3 patients. ||CI and p value could not be calculated because no case was missed by either CNN or radiologist.

Table 4: Sensitivity in test sets stratified by tumour stage and size
invariably lower the specificity. Given the somewhat low incidence of pancreatic cancer (38 per 100,000 population aged over 50 years), screening the population by a test with even 99% specificity would yield a positive predictive value of merely 3.6% (ie, only 3.6% of all individuals diagnosed as having pancreatic cancer actually have pancreatic cancer), causing excessive patient anxiety and unnecessary examinations and treatments.15 Instead of diagnosing whether the patient had pancreatic cancer (patient-level analysis) on the basis of a single round of analysis, in this study the images were processed into overlapping patches representing fine-grained sub-regions and analysed individually (patch-level analysis). Furthermore, because the patches were overlapping, each sub-region was analysed multiple times by the trained CNN, each time with different surrounding foci included for analysis, which thereby increased the chance of spotting foci harbouring cancer (ie, improved the sensitivity in detecting pancreatic cancer). However, just as multiple comparisons increase the chance of type I error and thus the standard for inferring statistical significance needs to be tightened, examining the image multiple times also increased the chance of false positivity (ie, reduced specificity); therefore, a cancer diagnosis should require more than one patch diagnosed as cancerous. In the patient-level analysis, diagnosing pancreatic cancer required the proportion of patches diagnosed as cancerous to exceed a threshold, which was objectively decided from the training and validation set, to protect against excessive false positives. Our results support the usefulness of this analytical approach.

To ascertain the potential generalisability of CNN in recognising pancreatic cancer, we tested the CNN trained solely using Asian patients with external datasets, which differed substantially in patient race and ethnicity and scanning parameters. Although the sensitivity for tumours smaller than 2 cm seemed unsatisfactory at first glance, DeWitt and colleagues17 found that the sensitivity of CT with radiologist interpretation was only 53% for pancreatic cancers smaller than 2.5 cm.19 Therefore, the sensitivity of the CNN-based analysis was at least similar to radiologist interpretation. The lower sensitivity of the CNN in the external test set compared with local test sets might be attributed to the differences in race and ethnicity and scanners and protocols between the training and external test sets, which could present greater challenges in small tumours. A key factor that influences the imaging characteristics of the pancreas and differs between races is the fat content of the pancreas. Higher pancreatic fat content reduces the density of the pancreas on CT (ie, darker pancreas to the naked eye), and previous studies found striking differences in pancreatic fat content between races and ethnicities.20,21 Although technical differences between different centres might have also contributed to the lower sensitivity of the CNN in the external test set, a standardised imaging protocol (the pancreatic protocol22) is widely adopted for the assessment of pancreatic diseases. Therefore, technical variations between institutions were more likely to have been random rather than institution-specific and less influential compared with differences in the imaging characteristics of the pancreas between races and ethnicities. Whether the performance of CNN can be improved by including in the training dataset images obtained in diverse patient populations using various scanners and protocols should be further investigated.

We used CT images with a multitude of technical variations as test sets to investigate whether CNN has the potential to distinguish patients with and without pancreatic cancer in real clinical practice. In a preliminary report, deep learning differentiated between 156 patients with pancreatic cancer and 300 healthy participants with 94.1% sensitivity and 98.5% specificity, but all images were obtained using a standardised protocol on CT scanners from a single vendor and carefully matched on the basis of imaging parameters to minimise technical variations across images.23 Therefore, the wide variations resulting from equipment, scanner, and various patient factors that radiologists inevitably encounter in clinical practice were excluded in that study, with unclear generalisability to other datasets. By contrast, this study included images obtained using six CT scanners from four major CT vendors, without selection based on imaging parameters. Therefore, the training and local test data in this study closely resembled that encountered in clinical practice and the results should have better generalisability than do those of the previous study. The excellent performance of our CNN in all local test sets supported the notion that a well-trained CNN based on adequate training data could reliably detect pancreatic cancer and serve as a computer-aided diagnostic tool in real-world clinical practice. On the other hand, the decreased accuracy of our CNN in an external test set suggested that important differences in CT images might exist between different races and ethnicities, and such diversity needs to be included in the training data for the trained CNN to have good generalisability across diverse populations.

The finding that the CNN-based analysis achieved a higher sensitivity than radiologists did should be interpreted with caution. Radiologists were provided with key clinical information and the reasons for requesting CT examinations from the clinicians when they interpreted the images, whereas the CNN was not provided with any information other than CT images. The trained CNN examined whether pancreatic cancer existed in individual focused areas (ie, patches) generated from the ROIs segmented by radiologists. Therefore, the major usefulness of the CNN was to assist radiologists in deciding whether a lesion or suspicious area in the pancreas harbours pancreatic cancer. One of the common clinical scenarios in which the CNN is especially useful is when patients present with obstructive jaundice, a typical sign of pancreatic cancer in the pancreatic head, but
the CT findings are negative or equivocal. In such circumstances occult pancreatic cancers should be highly suspected even if no apparent mass is noted on CT, given that approximately 40% of pancreatic cancers smaller than 2 cm are missed on CT because of inconspicuous borders with surrounding tissue.5,10 This study showed that a fine-grained examination of the pancreatic head by the CNN in those patients could facilitate detection of occult pancreatic cancers and supplement radiologist interpretation. Because the CNN examined only the CT patches without knowing other information, such as the number of patches generated from each patient or ROI and the patient or ROI from which a certain patch was generated, the high accuracy of the trained CNN was most likely due to an ability to recognise the imaging characteristics of pancreatic cancer. Collectively, our results suggest that the CNN could supplement radiologists to reduce the miss rates, rather than outperform or replace radiologists. Further prospective studies in real-world clinical settings are needed to draw definite conclusions on the relative performance of the CNN and radiologist interpretation and how the CNN should be integrated into clinical practice to maximise the benefits. Whether CNNs can distinguish between pancreatic cancer and other pancreatic pathologies such as pancreatitis and other pancreatic tumours must also be studied.

Through our explorative analysis, we found that 50×50 pixels, equivalent to 3.5×3.5 cm, was the optimal patch size for detecting pancreatic cancers on CT. The performance of CNN is influenced by the relative size of the patch compared with that of the ROI. In this study, the lower AUCs achieved by CNNs based on patches smaller than 50×50 pixels might be because the patches were too small to contain enough information on the relationship between the tumour and adjacent tissue, which is necessary for the CNN to learn and diagnose with precision. On the other hand, CNNs based on larger patches yielded lower AUCs with greater variability, probably due to an increased probability of introducing more noise with oversized patches. The optimal patch size can vary depending on the target conditions that are to be analysed by CNN and need to be explored individually.

This study had several strengths. In two local test datasets containing variations in imaging parameters seen in real-world clinical settings, the CNN could enable differentiation between individuals with and without pancreatic cancer, with accuracy consistently exceeding 98%, and had a lower miss rate than radiologist interpretation did. Furthermore, the CNN achieved acceptable performance in an external test dataset that differed in patient race and ethnicity with the training dataset, even for tumours smaller than 2 cm. We also did comprehensive exploratory analyses to identify the optimal patch size for analysing pancreatic imaging by CNN.

This study also had limitations. Because manual labelling of pancreatic images is extremely labour-intensive and access to external data is limited, we used a dataset of modest sample size that included only Asian participants from a single institution; therefore, the training dataset might not include the whole spectrum of the CT manifestations of pancreatic cancer. In response to this limitation, we used data augmentation techniques to increase the training data and verified the generalisability of the CNN using an external dataset.16 We seek to increase the volume and diversity of training and testing images in a future study. Also, because patients with pancreatic cancer and controls in the external test set were from different institutions, technical variations in image acquisition between institutions might be a potential source of confounding. Given the barriers limiting the export of patient data outside of hospitals and the paucity of publicly available data, we could not obtain data from a single North American source to constitute the external test set. However, the excellent performance of the CNN in the local test sets and the significant positive correlation between sensitivity and tumour size in the external test set suggested that the CNN diagnosed the presence of cancer on the basis of differential imaging characteristics rather than technical variations. Had the CNN relied on a subtle technical difference between the two sources in the external test set (eg, a mark placed on the image in one of the centres) for classification, the performance of the CNN would have been perfect and not related to the size of the tumour. Furthermore, because the mark was not present in the local training set, the CNN probably could not have learned to differentiate between pancreatic cancer and control on the basis of the presence or absence of the distinct technical difference.

In conclusion, this study provided a proof of concept that CNN can accurately distinguish pancreatic cancer on portal venous CT images. The CNN model holds promise as a computer-aided diagnostic tool to assist radiologists and clinicians in diagnosing pancreatic cancer.

Contributors
This study was conceived by K-LL, W-CL, and WW, with inputs from all other authors. K-LL and TW contributed equally in this work. K-LL, P-TC, and W-CL were involved in data acquisition. TW, YMT, HR, and WW did the analyses with critical inputs from all other authors. P-TC and TW were involved in data interpretation and the initial draft of the manuscript. K-LL, W-CL, WW, and M-SW were involved in data interpretation and critical revision of the manuscript. All authors approved the final version of the manuscript.

Declaration of interests
K-LL, TW, P-TC, YMT, M-SW, W-CL, and WW report grants from Taiwan Ministry of Science and Technology during the conduct of the study. K-LL, TW, P-TC, W-CL, and WW have a patent pending—differentiation between pancreatic cancer and non-cancerous pancreas on contrast-enhanced CT by deep learning. HR declares no competing interests.

Data sharing
The datasets from National Taiwan University Hospital were approved for the current study and are not publicly available. The Cancer Imaging Archive–Pancreas-CT dataset and Medical Segmentation Decathlon dataset are available online. The experimental and implementation methods are provided in detail in the appendix (pp 3–6). The major components and source codes of this study are available in the repositories Python, Keras, TensorFlow, and GitHub.
Acknowledgments
This study was supported by Ministry of Science and Technology (MOST; MOST 107-2611-F-002-014, MOST 107-2634-F-002-016, MOST 108-2634-F-002-011, MOST 108-2634-F-002-013, and MOST 109-2634-F-002-028), MOST All vista Healthcare Sub-center, and National Center for Theoretical Sciences Mathematics Division. The authors thank Nico Wu for assistance in data management and preprocessing, and Huifuxan Yen, Yenting Lin, and Po-chuan Wang for technical support.

References
1. Rahib L, Smith BD, Aizenberg R, Rosenzweig AE, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 2014; 74: 2913–20.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7–30.
3. Ryan DP, Hong TS, Bardeesy N. Pancreatic adenocarcinoma. N Engl J Med 2014; 371: 1039–49.
4. Al-Hwayry MM, Francis IR, Chari ST, et al. Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the Society of Abdominal Radiology and the American Pancreatic Association. Radiology 2014; 270: 248–50.
5. Dewitt J, Devereaux BM, Lehman GA, Sherman S, Imperiale TF. Comparison of endoscopic ultrasound and computed tomography for the preoperative evaluation of pancreatic cancer: a systematic review. Clin Gastroenterol Hepatol 2006; 4: 717–25; quiz 644.
6. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med 2018; 15: e2634002.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–18.
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402–10.
9. Yasaka K, Akai H, Abe O, Kiyu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 2018; 286: 887–96.
10. Frampas E, David A, Regenet N, Touchard F, Meyer J, Morla O. Pancreatic carcinoma: key points from diagnosis to treatment. Diagn Interv Imaging 2016; 97: 1207–23.
11. DeWitt J, Devereaux B, Chriswell M, et al. Comparison of endoscopic ultrasonography and multidetector computed tomography for detecting and staging pancreatic cancer. Ann Intern Med 2004; 141: 753–63.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2014; published online Sept 4. http://arxiv.org/abs/1409.1556 (preprint).
13. Panchapagesan S, Sun M, Khar A, et al. Multi-task learning and weighted cross-entropy for DNN-based keyword spotting. Interspeech; San Francisco, CA, USA; Sept 8–12, 2016 (abstr Fr-I-2-A-5).
14. Chollet F. Keras: the Python deep learning library. Astrophysics source code library. June 1, 2018. https://ui.adsabs.harvard.edu/abs/2018ascl.soft06023C (accessed Feb 26, 2020).
15. Dillon JV, Langmore I, Tran D, et al. TensorFlow distributions. arXiv 2017; published online Nov 28. http://arxiv.org/abs/1711.10604 (rnt). 16. Simpson AL, Antonelli M, Bakas S, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv 2019; published online Feb 25. http://arxiv.org/abs/1902.09063 (preprint).
17. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013; 26: 1045–57.
18. Berryman S, Pancreas-CT. The Cancer Imaging Archive (TCIA) Public Access. https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT (accessed Feb 26, 2020).
19. Roth HR, Lu L, Farag A, et al. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab N, Hornegger J, Wells W, Fragki A, eds. Medical image computing and computer-assisted intervention—MICCAI 2015. Lecture notes in computer science, vol 9349. Chiam, Switzerland: Springer, 2015: 556–64.
20. Chen W, Ji H, Feng J, et al. Classification of pancreatic cystic neoplasms based on modularity images. In: Shi Y, Srok HI, Liu M, eds. Machine learning in medical imaging. Cham, Switzerland: Springer, 2018: 161–69.
21. Zhou Y, Xie L, Fishman EK, Yuille AL. Deep supervision for pancreatic cyst segmentation in abdominal CT scans. In: Despavatz M, Maien HE, Franz A, Jannin P, Collins D, Duchesne S, eds. Medical image computing and computer assisted intervention—MICCAI 2017. Cham, Switzerland: Springer, 2017: 222–30.
22. Li J, Lu J, Liang P, et al. Differentiation of atypical pancreatic neuroendocrine tumors from pancreatic ductal adenocarcinomas: using whole-tumor CT texture analysis as quantitative biomarkers. Cancer Med 2018; 7: 4924–31.
23. Fu M, Wu W, Hong X, et al. Hierarchical combinatorial deep learning architecture for pancreas segmentation of medical computed tomography cancer images. BMC Syst Biol 2018; 12: 56.
24. Gibson E, Giganiti F, Hu Y, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans Med Imaging 2018; 37: 1822–34.
25. Roth HR, Lu L, Lay N, et al. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med Image Anal 2018; 45: 94–107.
26. Yu Q, Xie L, Wang Y, Zhou Y, Fishman EK, Yuille AL. Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. IEEE conference on computer vision and pattern recognition; Salt Lake City, UT, USA; June 18–23, 2018.
27. Chu LC, Park S, Kuwamoto S, et al. Application of deep learning to pancreatic cancer detection: lessons learned from our initial experience. J Am Coll Radiol 2019; 16: 1338–42.
28. Liu W, Anguelou D, Erhan D, et al. SSD: single shot multibox detector. arXiv 2015; published online Dec 8. http://arxiv.org/abs/1512.02232 (preprint).
29. Noz H, Hong S, Han B. Learning deconvolution network for semantic segmentation. IEEE conference on computer vision and pattern recognition; Santiago, Chile; Dec 11–18, 2017.
30. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chest-x-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE conference on computer vision and pattern recognition; Honolulu, HI, USA; July 21–26.
31. Wang D, Khodah A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. arXiv 2016; published online June 18. http://arxiv.org/abs/1606.05718 (preprint).
32. Maninis K-K, Pont-Tuset J, Arbelaez P, Van Gool L. Deep retinal image understanding. In: Ourselin S, Joskowicz L, Sabuncu M, et al. Medical image computing and computer-assisted intervention—MICCAI 2016. Cham, Switzerland: Springer, 2016: 140–48.
33. Agarwal B, Correa AM, Ho L. Survival in pancreatic carcinoma based on tumor size. Pancreas 2008; 36: e15–20.
34. Wu W, He X, Yang L, et al. Rising trends in pancreatic cancer incidence and mortality in 2000–2014. Clin Epidemiol 2018; 10: 789–97.
35. Chari ST. Detecting early pancreatic cancer: problems and prospects. Semin Oncol 2007; 34: 284–94.
36. Pannala R, Basu A, Petersen GM, Chari ST. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol 2009; 10: 88–95.
37. Szczepaniak LS, Victor RG, Mathur R, et al. Pancreatic steatosis and neuroendocrine tumors from pancreatic ductal adenocarcinomas: using whole-tumor CT texture analysis as quantitative biomarkers. Cancer Med 2018; 7: 4924–31.