RESEARCH ARTICLE

Pancreatic cancer grading in pathological images using deep learning convolutional neural networks [version 1; peer review: 1 approved, 1 approved with reservations]

Muhammad Nurmahir Mohamad Sehmi ¹, Mohammad Faizal Ahmad Fauzi ¹, Wan Siti Halimatul Munirah Wan Ahmad ¹, Elaine Wan Ling Chan²

¹Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

²International Medical University, Kuala Lumpur, Kuala Lumpur, Malaysia

V1 First published: 18 Oct 2021, 10:1057
<https://doi.org/10.12688/f1000research.73161.1>
 Latest published: 01 Nov 2022, 10:1057
<https://doi.org/10.12688/f1000research.73161.2>

Abstract

Background: Pancreatic cancer is one of the deadliest forms of cancer. The cancer grades define how aggressively the cancer will spread and give indication for doctors to make proper prognosis and treatment. The current method of pancreatic cancer grading, by means of manual examination of the cancerous tissue following a biopsy, is time consuming and often results in misdiagnosis and thus incorrect treatment. This paper presents an automated grading system for pancreatic cancer from pathology images developed by comparing deep learning models on two different pathological stains.

Methods: A transfer-learning technique was adopted by testing the method on 14 different ImageNet pre-trained models. The models were fine-tuned to be trained with our dataset.

Results: From the experiment, DenseNet models appeared to be the best at classifying the validation set with up to 95.61% accuracy in grading pancreatic cancer despite the small sample set.

Conclusions: To the best of our knowledge, this is the first work in grading pancreatic cancer based on pathology images. Previous works have either focused only on detection (benign or malignant), or on radiology images (computerized tomography [CT], magnetic resonance imaging [MRI] etc.). The proposed system can be very useful to pathologists in facilitating an automated or semi-automated cancer grading system, which can address the problems found in manual grading.

Keywords

digital pathology, pancreatic cancer, cancer grading, deep learning, image classification

Open Peer Review

Approval Status

	1	2
version 2 (revision) 01 Nov 2022	 view	
	↑	
version 1 18 Oct 2021	 view	 view

1. **Francisco Maria Calisto** , University of Lisbon, Lisbon, Portugal
University of Lisbon, Lisbon, Portugal
2. **Rajasvaran Logeswaran** , City University, Petaling Jaya, Malaysia

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research Synergy Foundation gateway**.

This article is included in the **Artificial Intelligence and Machine Learning gateway**.

Corresponding author: Mohammad Faizal Ahmad Fauzi (faizal1@mmu.edu.my)

Author roles: **Mohamad Sehmi MN:** Investigation, Software, Validation, Writing – Original Draft Preparation; **Ahmad Fauzi MF:** Methodology, Supervision; **Wan Ahmad WSHM:** Project Administration, Writing – Review & Editing; **Wan Ling Chan E:** Data Curation, Resources

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the internal grants from International Medical University [IMU 434/2019] and Multimedia University.

Copyright: © 2021 Mohamad Sehmi MN *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Mohamad Sehmi MN, Ahmad Fauzi MF, Wan Ahmad WSHM and Wan Ling Chan E. **Pancreatic cancer grading in pathological images using deep learning convolutional neural networks [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2021, **10**:1057 <https://doi.org/10.12688/f1000research.73161.1>

First published: 18 Oct 2021, **10**:1057 <https://doi.org/10.12688/f1000research.73161.1>

Introduction

Pancreatic cancer is one of the most lethal malignant neoplasms in the world,¹ developed when cells multiply and grow out of control in the pancreas,² forming cancer cells caused by cells mutation in their genes.³ Doctors commonly perform a biopsy to diagnose cancer when physical examination or imaging tests like magnetic resonance imaging (MRI) and computerized tomography (CT) scans are insufficient. In pancreatic cancer, grading is essential for planning treatment but is currently done using a meticulous microscopic examination.⁴ Up to now, there has been no successful implementation of artificial intelligence (AI) for classifying pancreatic cancer grade. The absence of such AI work motivates this paper to use transfer-learning to grade pathological pancreatic cancer images using 14 deep learning (DL) models. This work can facilitate an automated cancer grading system to address the exhaustive work of manual grading.

Pancreatic cancer and digital pathology

Pancreatic cancer is considered to be under-studied, and improvements in the diagnosis and prognosis of pancreatic cancer have therefore been minor.⁵ Digital pathology is an image-based environment obtained by scanning tissue samples from glass slides. Staining, usually using May-Grünwald-Giemsa (MGG) and haematoxylin and eosin (H&E) stains, is carried out on the tissue samples before digitization into whole-slide images. The cancer grade is identified by the degree of differentiation of the tumour cells⁶ ranging from well to poorly differentiated as described in [Table 1](#).

Deep learning and related works

Convolutional neural network (CNN) is a widely used deep learning (DL) algorithm in medical image-based classification and prediction.⁷ Several methods use CNN in cancer detection and diagnosis⁸ such as the Gleason grading of prostate cancer,^{9–11} colon cancer grading,¹² breast cancer detection,^{13,14} and pancreatic cancer detection^{15–18} and classification.¹⁹ However, grading of pancreatic cancer with DL still needs comprehensive study.

Methodology

The methodology of this work was done at Multimedia University, Cyberjaya, from June 2020 to May 2021. The overall methodology of this research is as illustrated in [Figure 1](#), with two major stages. In the data preparation stage, pathology images of pancreas tissue samples were obtained from our collaborator and the images were pre-classified by a pathologist into four classes. In the DL model development stage, the images were trained on the DL model and evaluated accordingly. All stages were carried out using [Jupyter notebook](#) in [Google Colab](#). The source code is available from [GitHub](#) and archived with Zenodo.²⁴

Table 1. Pancreatic cancer grade. MGG = May-Grünwald-Giemsa; H&E = haematoxylin and eosin.

Grade	Normal	Grade I	Grade II	Grade III
Description	Benign. Cells are not cancerous and will not spread.	Well differentiated. Cancer cells look like normal cell and are not growing rapidly.	Moderately differentiated. Cancer cells look abnormal and are growing faster than normal cell.	Poorly differentiated. Cancer cells look very abnormal and may spread aggressively.
MGG Stain
H&E Stain

Figure 1. Flowchart of the research work.

Ethical approval

This work was approved by the Research Ethics Committee of Multimedia University with approval number EA2102021. This article does not contain any studies with human participants or animals performed by any of the authors. Only pathology images were used, and the patients' personal data were anonymized.

Dataset preparation

Pathology image procurement

A total of 138 high-resolution images with varying dimensions (1600×1200 , 1807×835 and 1807×896) were obtained and pre-classified by the collaborators (see *Acknowledgements*). Four classes were identified (as shown in [Table 2](#)): Normal, Grade-I, Grade-II and Grade-III. Each image consisted of a tissue-sample stained with MGG and H&E. The image distribution in each class was unequal with Grade-II having 58 images and Normal with only 20 images. To better capture the cells characteristics which is paramount in determining their grade and to match the lower-resolution setting of the network's input, the images were pre-processed into small non-overlapping patches.

Image pre-processing

The pre-trained models require a low dimension and square image for training and making predictions. The squared slicing method is used where smaller patches with approximately 200×200 pixels of non-overlapping regions are sampled from the original images. Further processing was done to remove unwanted patches, as shown in [Figure 2](#).

Image dataset

A total of 6468 patches were generated from the slicing process of the 138 original images which is an increase of 468% in number of images. Overall, 50.5% (3267) of the patches with background and non-tissue information were discarded, and the remaining are listed in [Table 3](#). Examples of MGG stain and H&E stain pathology images are shown in [Table 1](#), with the mixed dataset combining all images from MGG and H&E stains. From the numbers in [Table 3](#), these datasets still had an imbalanced number of patch images in each class but this can be mitigated by employing a weighted average to evaluate the model.

Training-validation splitting and K-fold cross-validation

To evaluate the DL model, images in each dataset were split into training and validation set with 80-20 ratio. K-fold cross-validation with $K = 5$ was used by splitting all MGG, H&E and the mixed dataset into five parts, producing

Table 2. Number of high-resolution images in the dataset. MGG = May-Grünwald-Giemsa; H&E = haematoxylin and eosin.

Stain\Class	Normal	Grade I	Grade II	Grade III	Total
MGG stained	13	4	43	19	79
H&E stained	7	27	15	10	59
Total	20	31	58	29	138

Figure 2. Process of slicing an image and discarding unwanted non-tissue patches.

Table 3. Number of sliced images kept for training and validation. MGG = May-Grünwald-Giemsa; H&E = haematoxylin and eosin.

Stain\Class	Normal	Grade I	Grade II	Grade III	Total
MGG stained	401	108	983	366	1858
H&E stained	139	606	309	289	1343
Total	540	714	1292	655	3201

cross-validation sets with five new copies of MGG, H&E and mixed datasets and labelled (e.g. MGG Set 1 to MGG Set 5 for MGG). Each set had a different set of images used for training (80%) and validation (20%). The average value was calculated from the five-training iterations to evaluate the performance.

Image data augmentation and normalisation

Image data augmentation was implemented to virtually expand the training set, but not on the validation set. The transformation parameter involved was horizontal flip, vertical flip and -90° to 90° rotation range. Image data normalisation was used to rescale image pixels from a range of [0,255] to [0,1] so the input pixels will have similar data distribution.

CNN deep learning model development

Deep CNN algorithm was used for developing a model for classifying pancreatic cancer grading from pathology images.

Table 4. ImageNet pre-trained models.

Pre-trained models	Input shape	Top-1 accuracy
Xception	299 × 299	0.790
VGG16	224 × 224	0.713
VGG19	224 × 224	0.713
ResNet50V2	224 × 224	0.760
ResNet101V2	224 × 224	0.772
ResNet152V2	224 × 224	0.780
InceptionV3	299 × 299	0.779
InceptionResNetV2	299 × 299	0.803
MobileNetV2	224 × 224	0.713
DenseNet121	224 × 224	0.750
DenseNet169	224 × 224	0.762
DenseNet201	224 × 224	0.773
NASNetMobile	224 × 224	0.744
NASNetLarge	331 × 331	0.825

Transfer-learning

A total of 14 CNN pre-trained models from recognizing 1000 classes in ImageNet was selected from Keras API²⁰ to get the best model for classifying the 4-grade classes of pancreatic cancer. The proposed pre-trained models are listed in Table 4 along with their original model's image input shape, and its respective top-1 accuracy on the ImageNet validation set.

Fine-tuning

All 14 models were fine-tuned with four newly added layers to extract the features from pathology images: a flatten-layer to form a 1D fully connected layer; a dense-layer with 256 nodes and ReLu activation-function; a dropout-layer with a rate of 0.4 to regularise the network; and lastly another dense-layer with 4-nodes and SoftMax activation-function to normalize the probability of prediction.

Setup and evaluation parameters

Batch size of 64 was chosen to allow the computer to train and validate 64-patch-samples at the same time. Adam optimizer with default initial learning rate of $\alpha = 0.01$ and moment decay rate of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used. The loss function is calculated using categorical cross-entropy for the 4-class classification task. With this setup, the models are compiled and trained for 100-epochs.

The confusion matrix, precision, recall, f1-score and weighted-average were used to evaluate the model's performance. Weighted-average was used to calculate the performance of individual cross-validation set and suitable for imbalanced dataset. The equation for the weighted-average is:

$$\text{average}_{(\text{weighted})} = \sum_{k=1}^n \left(P_k \frac{\text{No. of images in class } k}{\text{Total number of images in dataset}} \right)$$

Results and discussion

Effect of data augmentation

This experiment was done with the first cross-validation set of the mixed dataset, to observe how data augmentation impacts a model training performance.²³ Table 5 and Table 6 display the final accuracy and loss of training and validation set after 100 epochs. Without data augmentation in Table 5, it is evident that overfitting has occurred, because the model is doing very well on the training set but not on the validation set. With data augmentation, the validation accuracy

Table 5. Model accuracy for without and with data augmentation after 100 epochs.

Pre-trained models	Without image data augmentation		With image data augmentation	
	Training set (%)	Validation set (%)	Training set (%)	Validation set (%)
Xception	99.88	83.75	93.63	85.02
VGG16	97.35	76.88	86.48	81.12
VGG19	64.85	54.83	77.34	77.22
ResNet50V2	100.00	82.66	95.39	86.11
ResNet101V2	99.92	79.84	93.44	85.80
ResNet152V2	100.00	81.25	95.35	86.58
InceptionV3	99.38	82.65	91.23	83.15
InceptionRes-NetV2	99.61	79.06	90.12	83.78
MobileNetV2	99.68	82.50	94.14	85.02
DenseNet121	99.84	82.81	94.69	88.14
DenseNet169	99.84	85.00	95.98	89.70
DenseNet201	99.92	85.47	96.76	88.14
NASNetMobile	99.92	79.69	91.21	81.75
NASNetLarge	98.32	73.48	89.84	79.88

Table 6. Model loss for without and with data augmentation after 100 epochs.

Pre-trained models	Without image data augmentation		With image data augmentation	
	Training set	Validation set	Training set	Validation set
Xception	0.00241	1.49118	0.16254	0.47146
VGG16	0.09256	0.87281	0.35639	0.46890
VGG19	0.95239	1.14815	0.56213	0.60279
ResNet50V2	0.00005	1.40010	0.12776	0.42690
ResNet101V2	0.00287	1.31044	0.17701	0.41832
ResNet152V2	0.00010	1.68771	0.12355	0.40160
InceptionV3	0.01608	1.02646	0.23775	0.44671
InceptionRes-NetV2	0.02074	1.15162	0.26173	0.47013
MobileNetV2	0.00703	1.03879	0.14739	0.47216
DenseNet121	0.00615	0.92605	0.14379	0.33270
DenseNet169	0.00423	0.93888	0.11209	0.35975
DenseNet201	0.00385	0.94979	0.09399	0.32288
NASNetMobile	0.00338	1.80272	0.22978	0.47335
NASNetLarge	0.05669	3.36376	0.32225	0.68587

improved, specifically on VGG19 model (54.83% to 77.22%). The training accuracy of other models are slightly reduced with data augmentation (except for VGG19) but it is normal as the model is learning newly transformed images. The validation loss also shows a reduction, as in [Table 6](#), such as on NASNetLarge model from 3.36376 to 0.68587. Overall, these results show that data augmentation may reduce overfitting and improve model performance as reported in.^{10,13,14}

Comparison analysis of model performance

The overall performance results of all 14 different transfer-learning models proposed for this experiment are presented. Each model was trained with the 3 datasets and 5-fold cross-validation. [Figure 3](#) illustrates the overall performance in terms of mean f1-score.

Figure 3. Mean F1-score of models.

Comparison between MGG, H&E and the mixed dataset

This comparison shows how a DL model learns from single coloured stain. In [Figure 3](#), all models trained with the H&E obtained the highest f1-score compared to MGG and mixed. Most models scored above 0.9 except for VGG19 (0.87). When trained with the MGG, models other than VGG16 and VGG19 performed the lowest compared to H&E and Mixed. The performance of mixed is as expected because it contains a mixture of both datasets. The VGG16 and VGG19 model, however, performed better on MGG than Mixed, due to the small VGG network architecture and small fully-connected layers making it unable to learn complex features and patterns in pathology image. The trend described in [Figure 3](#) indicates that image patches in the H&E are easier to learn with better prediction than MGG.

Comparison between pre-trained models

From the result, DenseNet network architecture was the best at classifying pathology images where all three variations trained on MGG, H&E and mixed take the top spot among the 14 models. The ResNet on mixed dataset were ranked ascendingly from ResNet101V2, ResNet50V2 and ResNet152V2 before the three DenseNet models. This supports the work of Huang *et al.* in²¹ where DenseNet was designed to improve the ResNet architecture. DenseNet201 which is a much deeper layer than the other two DenseNet models managed to achieve the highest f1-score of 0.88, 0.96 and 0.89 for MGG, H&E and mixed, respectively. The DenseNet121 and DenseNet169 performance on the three dataset scores were marginally lower at 0.87, 0.95, 0.89 and 0.87, 0.95, 0.88, respectively. This shows that a deeper DenseNet layer can perform more accurate prediction.

The Xception²¹ and InceptionResNetV2²² are improvements of InceptionV3, which performs better than their ancestor. The f1-scores for Xception trained by MGG, H&E and Mixed are 0.85, 0.94 and 0.86, as compared to InceptionV3, 0.80, 0.92 and 0.83, respectively. However, InceptionResNetV2 are just slightly higher than InceptionV3 (0.93 and 0.83 for H&E and mixed) but lower for MGG (0.80). VGG models did not perform quite well when compared to its earlier models. VGG19, which is supposed to be an improvement to VGG16, failed to achieve a greater f1-score, with 0.74, 0.87 and 0.65 for the MGG, H&E and mixed, respectively, while VGG16 was higher at 0.80, 0.93 and 0.78. The results concluded that VGG19 was the worst performing model for our datasets.

This experiment applied transfer-learning on 14 ImageNet pre-trained models to classify pancreatic cancer grades. From the comparisons, DenseNet201 model is suggested for practical application of pancreatic grading system of MGG or H&E stains.

Comparison between the best and the worst performing model

[Table 7](#) and [Table 8](#) shows the precision and recall of VGG19 (the worst) and DenseNet201 (the best) for the three datasets. VGG19 struggles to make prediction for Grade-I patches in MGG where the precision and recall are 0.00 for CV sets 3, 4, and 5. A similar pattern is noticed in Grade-III patches, and from our observation, this is because most of the Grade-I and Grade-III patches were wrongly predicted as Grade-II. This is due to the imbalance classes in MGG where

Table 7. Precision rate of VGG19 and DenseNet201.

Precision		VGG19					DenseNet201				
Class\CV set		1	2	3	4	5	1	2	3	4	5
MGG Dataset	Normal	0.87	0.91	0.96	0.89	0.85	0.95	0.93	0.99	0.94	0.95
	Grade I	0.62	0.64	0.00	0.00	0.00	0.77	0.85	1.00	0.79	0.70
	Grade II	0.77	0.80	0.74	0.72	0.77	0.87	0.87	0.91	0.86	0.90
	Grade III	0.71	0.56	0.54	0.56	0.64	0.79	0.83	0.80	0.85	0.84
	Weighted Average	0.77	0.77	0.70	0.68	0.72	0.87	0.87	0.91	0.87	0.89
	Mean	0.7289					0.8819				
H&E Dataset	Normal	1.00	0.96	0.96	1.00	0.96	1.00	1.00	1.00	1.00	1.00
	Grade I	0.98	0.95	0.97	0.91	0.91	0.98	1.00	0.99	0.97	0.97
	Grade II	0.92	0.89	0.84	0.90	0.86	0.93	0.94	0.87	0.95	0.86
	Grade III	0.95	0.75	0.81	0.90	0.93	0.92	0.93	0.88	0.98	0.95
	Weighted Average	0.89	0.87	0.89	0.88	0.88	0.96	0.97	0.94	0.97	0.94
	Mean	0.8802					0.9565				
Mixed Dataset	Normal	0.90	0.86	0.69	0.76	0.61	0.94	0.88	0.89	0.96	0.94
	Grade I	0.95	0.95	0.90	0.81	0.87	0.93	0.97	0.98	0.96	0.95
	Grade II	0.98	0.67	0.65	0.52	0.56	0.85	0.85	0.89	0.88	0.87
	Grade III	0.78	0.88	0.88	0.00	0.00	0.83	0.86	0.82	0.91	0.83
	Weighted Average	0.92	0.81	0.76	0.52	0.52	0.88	0.88	0.90	0.92	0.89
	Mean	0.7055					0.8935				

Table 8. Recall Rate of VGG19 and DenseNet201.

Recall		VGG19					DenseNet201				
Class\CV set		1	2	3	4	5	1	2	3	4	5
MGG Dataset	Normal	0.89	0.93	0.90	0.84	0.88	0.89	0.97	0.93	0.94	0.94
	Grade I	0.23	0.41	0.00	0.00	0.00	0.45	0.50	0.55	0.52	0.76
	Grade II	0.92	0.90	0.93	0.94	0.93	0.94	0.94	0.95	0.94	0.94
	Grade III	0.47	0.42	0.37	0.30	0.44	0.77	0.71	0.88	0.73	0.71
	Weighted Average	0.78	0.78	0.76	0.74	0.77	0.87	0.88	0.91	0.87	0.88
	Mean	0.7669					0.8819				
H&E Dataset	Normal	1.00	0.96	0.96	0.93	0.82	1.00	1.00	1.00	1.00	1.00
	Grade I	0.80	0.91	0.93	0.91	0.93	0.99	0.99	0.98	0.99	0.97
	Grade II	0.92	0.90	0.92	0.97	0.87	0.87	0.94	0.89	0.95	0.92
	Grade III	0.72	0.83	0.76	0.76	0.72	0.95	0.95	0.90	0.93	0.90
	Weighted Average	0.84	0.87	0.88	0.87	0.86	0.95	0.97	0.94	0.97	0.95
	Mean	0.8651					0.9571				
Mixed Dataset	Normal	0.76	0.81	0.92	0.29	0.57	0.94	0.98	0.97	0.98	0.94
	Grade I	0.75	0.78	0.74	0.79	0.78	0.90	0.87	0.89	0.92	0.85
	Grade II	0.95	0.95	0.92	0.93	0.90	0.90	0.92	0.91	0.94	0.92
	Grade III	0.46	0.37	0.11	0.00	0.00	0.78	0.75	0.82	0.81	0.82
	Weighted Average	0.77	0.77	0.71	0.60	0.63	0.88	0.88	0.90	0.92	0.89
	Mean	0.6981					0.8933				

Grade-II patches account for 52.9% of the total images whereas Grade-I consist of only 5% and Grade-III 19.7%. This class imbalance has caused the VGG19 model to struggle a lot at recalling class with fewer data.

For H&E images, the effect of class imbalance however did not affect the performance of VGG19. The recall and precision for Normal class are ranked among the highest despite its smallest number (10%) of patches. Looking back at [Table 1](#), the H&E Normal images have a quite different stain colour compared to other classes, which explains the good prediction for both models. This could be seen as a problem where limited image variation can cause biasness. The precision of Normal class would score poorly if it were tested to predict different variation of H&E stain image even with the same set of ground truth, but can be assuaged if the class have many different variations of stain colour.

For the mixed dataset, VGG19 also struggled to predict Grade-III class, especially on CV 4 and 5 where it scored 0.00 for both metrics. The reason could be that the Grade-III patches are difficult for the VGG19 model to learn. This is the reason why cross-validation should be performed to rigorously evaluate a DL model. DenseNet201 managed to get good recall for Grade-III patches for both CV sets, confirming its ability to learn complex features on the pathology image.

Conclusion

This paper presents development of several deep learning models through transfer-learning for classifying pancreatic cancer grade from pathology images. The datasets were trained on a total of 14 ImageNet pre-trained models. Image data augmentation was performed to counter the low number of images and has proven to improve the validation accuracies of all pre-trained models up to 40%. The evaluation on 14 pre-trained models shows that the DenseNet models performed best compared to the other models. Most of the models trained by H&E managed to achieve f1-score above 0.9. The MGG dataset scores lower f1-score compared to the mixed dataset. The highest f1-scores were achieved by DenseNet201, with 0.8786, 0.9561 and 0.8915 for MGG, H&E and Mixed, respectively. To the best of our knowledge, no similar work on pancreatic cancer grading has been reported in the literature. With these promising early results, this work can aid pathologists in facilitating an automated pancreatic cancer grading system for better cancer diagnosis and prognosis. This study has not been tested with whole slide images (WSI), but similar approaches can be applied. Further improvements to the system can potentially be achieved by using future state-of-the-art DL models.

Data availability

Underlying data

Open Science Framework: Dataset for Pancreatic Cancer Grading in Pathological Images using Deep Learning Convolutional Neural Networks. <https://doi.org/10.17605/OSF.IO/WC4U9>.²³

This project contains the following underlying data:

- Dataset PCGIPI-Original.zip (pancreatic pathological image patches used for our analysis. The stain types are May-Grünwald-Giemsa (MGG) and Haematoxylin and Eosin (H&E)).
- Dataset PCGIPI-sliced.zip
- PCGIPI Results.xlsx
- Slicing Process for Table 3.docx

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Extended data

Analysis code available from: <https://github.com/mnmahir/FYProject-PCGIPI>

Archived analysis code as at time of publication: <https://doi.org/10.5281/zenodo.5532663>.²⁴

License: [MIT](#)

Acknowledgements

We would like to thank our collaborators Clinipath (Malaysia) Sdn. Bhd. for providing the image dataset and their ground truth for evaluation.

References

1. Rawla P, Sunkara T, Gaduputi V: **Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors.** *World J. Onco.* 2019; **10**(1): 10–27.
[Publisher Full Text](#) | [Free Full Text](#) | [PubMed Abstract](#)
2. A. C. Society: *What Is a Pancreatic Neuroendocrine Tumor?*. 2020, 17th September.
[Reference Source](#)
3. A. C. Society: *Changes in genes.* 2014, 17th September.
[Reference Source](#)
4. Haeberle L, Esposito I: **Pathology of pancreatic cancer.** *Translational gastroenterology and hepatology.* 2019; **4**: 50–50.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. McGuigan A, Kelly P, Turkington RC, et al.: **Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes.** *World J. Gastroenterol.* Nov 21 2018; **24**(43): 4846–4861.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Wasif N, et al.: **Impact of tumor grade on prognosis in pancreatic cancer: should we include grade in AJCC staging?**. *Ann. Surg. Oncol.* Sep 2010; **17**(9): 2312–20.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Latif J, Xiao C, Imran A, et al.: **Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review.** *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).* 2019. pp. 1–5.
8. Acs B, Rantalainen M, Hartman J: **Artificial intelligence as the next step towards precision pathology.** *J. Intern. Med.* Jul 2020; **288**(1): 62–81.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Arvaniti E, et al.: **Automated Gleason grading of prostate cancer tissue microarrays via deep learning.** *Sci. Rep.* Aug 13 2018; **8**(1): 12054.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Karimi D, Nir G, Fazli L, et al.: **Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images-Role of Multiscale Decision Aggregation and Data Augmentation.** *IEEE J. Biomed. Health Inform.* May 2020; **24**(5): 1413–1426.
[Publisher Full Text](#)
11. Li Y, et al.: **Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer.** *IEEE Access.* 2020; **8**: 117714–117725.
[Publisher Full Text](#)
12. Vuong TLT, Lee D, Kwak JT, et al.: **Multi-task Deep Learning for Colon Cancer Grading.** *2020 International Conference on Electronics, Information, and Communication (ICEIC).* 2020; pp. 1–2.
13. Alom MZ, Yakopcic C, Nasrin MS, et al.: **Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network.** *J. Digit. Imaging.* 2019/08/01 2019; **32**(4): 605–617.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Shahidi F, Daud SM, Abas H, et al.: **Breast Cancer Classification Using Deep Learning Approaches and Histopathology Image: A Comparison Study.** *IEEE Access.* 2020; **8**: 187531–187552.
[Publisher Full Text](#)
15. Chu LC, Fishman EK: **Deep learning for pancreatic cancer detection: current challenges and future strategies.** *The Lancet Digital Health.* 2020; **2**(6): e271–e272.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Chu LC, Park S, Kawamoto S, et al.: **Pancreatic Cancer Imaging: A New Look at an Old Problem.** *Curr. Probl. Diagn. Radiol.* 2021/07/01/2021; **50**(4): 540–550.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Liu KL, et al.: **Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation.** *Lancet Digit Health.* Jun 2020; **2**(6): e303–e313.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Sekaran K, Chandana P, Krishna NM, et al.: **Deep learning convolutional neural network (CNN) With Gaussian mixture model for predicting pancreatic cancer.** *Multimed. Tools Appl.* 2020/04/01 2020; **79**(15): 10233–10247.
[Publisher Full Text](#)
19. Naito Y, et al.: **A deep learning model to detect pancreatic ductal adenocarcinoma on endoscopic ultrasound-guided fine-needle biopsy.** *Sci. Rep.* 2021/04/19 2021; **11**(1): 8454.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Keras: *Keras Applications.* 15th December.
[Reference Source](#)
21. Chollet F: **Xception: Deep learning with depthwise separable convolutions.** *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017; pp. 1251–1258.
22. Szegedy C, Ioffe S, Vanhoucke V, et al.: **Inception-v4, inception-resnet and the impact of residual connections on learning.** *Thirty-first AAAI conference on artificial intelligence.* 2017.
23. Ahmad WSHMW, Sehmi MNM, Fauzi MFA, et al.: **Dataset for Pancreatic Cancer Grading in Pathological Images using Deep Learning Convolutional Neural Networks.** 2021, October 5.
[Publisher Full Text](#)
24. Sehmi M: **mmmahir/FYProject-PCGIPI: First release (v1.0.0).** *Zenodo.* 2021.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 19 July 2022

<https://doi.org/10.5256/f1000research.76792.r142271>

© 2022 Logeswaran R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rajasvaran Logeswaran

Faculty of Information Technology, City University, Petaling Jaya, Malaysia

Summary:

The paper reports on novel work undertaken in automated grading of pancreatic cancer using pathology images and deep learning. The classification is based on the grade of the cancer, instead of just detecting the presence or type of cancer, and uses pathology images as opposed to the conventional radiology images prevalent in most prior research efforts. The main advantage of the proposed system is that it overcomes the very tedious and error prone conventional pancreatic cancer grading method that requires manual microscopic examination of the cancerous tissues after biopsy. The reduced misdiagnosis promotes providing correct treatment planning for patients. Although using only a limited size training set, the proposed system was tested using 14 different ImageNet pre-trained models and achieved high accuracies of up to 95.61% for two different pathological stains.

Comments:

The paper is well written and the flow is both logical and guides the reader through the work in an easy to understand manner. The technical content is good with appropriate use of vocabulary and terminology. The subject matter is of current interest and the publication is a worthwhile contribution to the body of knowledge.

Recommendations for Improvement:

Some points of consideration for improvement of the paper are listed below:

1. *Abstract - Results*: It would be advisable to change "small sample set" to "small training set", to better represent the effectiveness even when training was limited. This is especially so as deep learning is employed for the grading system.
2. *Introduction*: Consider changing "paper" in "The absence of such AI work motivates this paper to" to "project" or "research".
3. *References*: Although the references are fairly recent, the paper would benefit by additionally citing the most recent relevant publications from the current year. It should also be checked

if "et al." can be used in the References or the names of all authors should be listed in the References section.

4. *Methodology*: In terms of language, better to change "The methodology of this work was done at" to "The research work was undertaken at".
5. *Figure 1*: Do use consistent style of verbs in the flowchart, e.g. change "Fine tuning" to "Fine tune" and "Training" to "Train".
6. *Pathology image procurement*: Consider changing "which is" to "that are".
7. *Image dataset*: Consider adding a comma before "which" in "images which is".
8. *Image data augmentation and normalization*: It is stated that the images were rescaled from [0,255] to [0,1]. However, was this as a grayscale intensity image or for each of the RGB (or other colour space) components? Do clarify as there are no further images shown, making it difficult to infer.
9. Equation numbering could be added.
10. *Effect of data augmentation*: Citation 23 appears to be out of place. Consider changing "is doing very well on the training set" to "performed well on the training set". The citations at the end of the last sentence should be moved to before the full-stop.
11. *Table 5*: Consider simplifying the caption to "Model accuracy after 100 epochs" or removing "for". Similarly for the *Table 6* caption.
12. *Figure 3*: The names of the dataset do not tally with the description given earlier in the text. For consistency, "Blue" and "Purple" should be changed to "MGG" and "H&E", respectively. The y-axis of the graph should be labelled ("Mean f1-score"). Also ensure consistency of spelling for "f1-score" in the caption and text. Similarly, ensure consistency of the spelling "Mixed" dataset (cf. "mixed").
13. *Comparison between pre-trained models*: Consider deleting ", which performs better than their ancestor."
14. *Comparison between the best and the worst performing model*: Consider changing "imbalance classes" to "imbalanced classes". Also, "recalling class with fewer data" to "recalling classes with less data".

Data availability:

The authors have made their data and results available for other researchers. This is useful as it allows for verification of the findings, should the need be, and can propagate further research in this area by helping overcome the scarcity of labelled data. The code used is also provided to assist other researchers.

Conclusion:

The paper presents useful findings that would be beneficial in quick and accurate pancreatic

cancer grading for correct timely treatment planning. The finding may also open avenues of research and implementation in for other types of cancers and tissue analysis. This work is a useful contribution to the body of knowledge in this domain.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical image processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 18 July 2022

<https://doi.org/10.5256/f1000research.76792.r142618>

© 2022 Calisto F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Francisco Maria Calisto

¹ Institute for Systems and Robotics (ISR/IST), LARSyS, Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal

² Interactive Technologies Institute (ITI), LARSyS, Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal

In this manuscript, the authors are presenting the development of several DL (deep learning) models through transfer-learning for classifying pancreatic cancer grade from pathology images. Specifically, the authors performed data augmentation across a dataset of images to counter the

low number of images and improve the validation accuracies of all pre-trained models. Overall, this is an exciting line of work, but the current draft of this manuscript has some minor concerns that must be addressed first before publication. Contributions, if so, must be described clearly. Additionally, the findings are also important attributes of such significant research work. As follows, further comments are detailing improvements for the next iterations of the manuscript.

Strengths

- 1.1.** The manuscript addresses key concerns about transfer-learning for classifying pancreatic cancer;
- 1.2.** The manuscript brings forward novel perspectives for DL developments in a specific medical domain;

Weaknesses

- 2.1.** Contributions and impact for this scientific community are not clear;
- 2.2.** Findings and implications of the work are merely discussed;

Requested Changes

There is much to appreciate in this manuscript, as it investigates several DL models for classifying pancreatic cancer. However, the provided findings and work implications are not enough, and the discussion section would need to be considerably strengthened by a more robust engagement with the cited literature. I would encourage the authors to think about how their findings extend or add to both DL and medical literatures in this space in order to refine their contribution to both fields. For instance, it would be interesting to understand the application of these DL models to a real scenario [1, 4]. At least, a small discussion on how could these techniques help real problems, such as different medical domains [2, 3, 5].

The paper needs a brief proofreading. Be aware of tenses also. It would be wise to proofread the paper and correct these, as well as other spelling errors.

The related work seems scarce for a work around the healthcare domain. The paper must discuss recent work in a methodological sense and base their work on the proposed approach. A lot of work is already available into this topic [1, 3, 5]. Specifically, the scientific community has numerous works on this topic [1, 5, 7], but authors could also map some literature around other domains [5, 6, 8], where these level of approaches are essential.

To conclude, the paper is almost ready for indexing and the authors did a splendid work. However, I strongly recommend the authors improve the manuscript for the next iterations of the work, as this is a chief importance domain.

Missing References

[1] <https://doi.org/10.1259/bjr.20220072>

[2] <https://doi.org/10.1016/j.artmed.2022.102285>

[3] <https://doi.org/10.1007/s00521-022-07183-8>

[4] <https://doi.org/10.1016/j.ijhcs.2021.102607>

[5] https://doi.org/10.1007/978-981-15-9735-0_2

[6] <https://doi.org/10.1145/3399715.3399744>

[7] <https://doi.org/10.1371/journal.pone.0195621>

[8] <https://doi.org/10.1145/3132272.3134111>

References

1. Marti-Bonmati L, Cerdá-Alberich L, Pérez-Girbés A, Díaz Beveridge R, et al.: Pancreatic cancer, radiomics and artificial intelligence. *Br J Radiol*. 2022. 20220072 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Calisto FM, Santiago C, Nunes N, Nascimento JC: BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artif Intell Med*. **127**: 102285 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Budak C, Mençik V: Detection of ring cell cancer in histopathological images with region of interest determined by SLIC superpixels method. *Neural Computing and Applications*. 2022. [Publisher Full Text](#)
4. Calisto F, Santiago C, Nunes N, Nascimento J: Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*. 2021; **150**. [Publisher Full Text](#)
5. Shinde S, Kulkarni U, Mane D, Sapkal A: Deep Learning-Based Medical Image Analysis Using Transfer Learning. **932**: 19-42 [Publisher Full Text](#)
6. Calisto FM, Nunes N, Nascimento JC: BreastScreening: On the Use of Multi-Modality in Medical Imaging Diagnosis. *AVI '20: Proceedings of the International Conference on Advanced Visual Interfaces*. 2020. [Publisher Full Text](#) | [Reference Source](#)
7. Niazi MKK, Tavorara TE, Arole V, Hartman DJ, et al.: Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning. *PLoS One*. 2018; **13** (4): e0195621 [PubMed Abstract](#) | [Publisher Full Text](#)
8. Calisto FM, Ferreria A, Nascimento JC, Gonçalves D: Towards Touch-Based Medical Image Diagnosis Annotation. *ISS '17: Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. 2020. [Publisher Full Text](#) | [Reference Source](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human-Computer Interaction, Health Informatics, Artificial Intelligence

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Sep 2022

Wan Siti Halimatul Munirah Wan Ahmad

Thank you Prof. F.M.Calisto for the encouraging comments for improving our article.

We have noted the following changes (new texts are emphasized in Bold):

Revised Introduction:

Pancreatic cancer is one of the most lethal malignant neoplasms in the world,¹ developed when cells multiply and grow out of control in the pancreas,² forming cancer cells caused by cells mutation in their genes.³ Doctors commonly perform a biopsy to diagnose cancer when physical examination or imaging tests like magnetic resonance imaging (MRI) and computerized tomography (CT) scans are insufficient. In pancreatic cancer, grading is essential for planning treatment but is currently done using a meticulous microscopic examination.⁴ **Limited work found on analysis of pathological images for pancreatic cancer. Niazi et al. [Suggested Reference 7] presented a deep learning method to differentiate between pancreatic neuroendocrine tumor and non-tumor regions based on Ki67 stained biopsies. The purpose is for the quantification of positive tumor cells in a hotspot.** Up to now, there has been no successful implementation of artificial intelligence (AI) for classifying pancreatic cancer grade. The absence of such AI work motivates this paper to use transfer-learning to grade pathological pancreatic cancer images using 14 deep learning (DL) models. This work can facilitate an automated cancer grading system to address the exhaustive work of manual grading.

Revised Deep learning and related works section:

Convolutional neural network (CNN) is a widely used deep learning (DL) algorithm in medical image-based classification and prediction.⁷ Several methods use CNN in cancer detection and diagnosis⁸ such as the Gleason grading of prostate cancer,^{9–11} colon cancer grading,¹² breast cancer detection,^{13,14} and pancreatic cancer detection^{15–18} and classification.¹⁹ **AI has been proven to assist clinicians with better prediction and faster diagnosis for breast cancer screening [Suggested Reference 2].** However, the

grading of pancreatic cancer with DL still needs comprehensive study.

New subsection at the end of Introduction:

Contributions:

This work presents an automated grading system focusing on pancreatic cancer from pathology images, which has not been done before to the best of our knowledge. The work also contributes a comparison of performance for 14 DL models on two different pathological stains, namely the May-Grünwald-Giemsa and haematoxylin and eosin.

Revised Effect of data augmentation section:

This experiment was done with the first cross-validation set of the mixed dataset, to observe how data augmentation impacts a model training performance.²³ Table 5 and Table 6 display the final accuracy and loss of training and validation set after 100 epochs. Without data augmentation in Table 5, it is evident that overfitting has occurred, because the model is doing very well on the training set but not on the validation set. With data augmentation, the validation accuracy improved, specifically on VGG19 model (54.83% to 77.22%). The training accuracies of other models are slightly reduced with data augmentation (except for VGG19) but it is normal as the model is learning newly transformed images. The validation loss also shows a reduction, as in Table 6, such as on NASNetLarge model from 3.36376 to 0.68587. Overall, these results show that data augmentation may reduce overfitting and improve model performance as reported in.^{10,13,14}**The reason behind this is the model is becoming more robust with data augmentation for getting to learn various transformed images of the limited size dataset. These kind of images are high-likely to exist in real-world applications, especially when it comes to unique human cells.**

Added last paragraph before Conclusion section:

From the study, we can see that integrating AI into the diagnosis system can assist the pathologist in getting the suggestive grading based on the prediction. It is however meant to assist and not on decision-making. The future aim of this study is to have a platform for screening of pancreatic cancer biopsies.

We look forward to getting this article indexed. Thank you very much.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research