Pancreatic Tumor Classification

Gary Gao
Cornell University
gg392@cornell.edu

Max Zhou
Cornell University
mz282@cornell.edu

Rohit Bandaru
Cornell University
rb696@cornell.edu

Seung Won Yoo
Cornell University
sy536@cornell.edu

Abstract

Pancreatic cysts are found in over 2% of abdominal CT scans [2], and may be either benign or malignant. Early detection and classification of these cysts is important both for ensuring prompt treatment and survival in case of malignant or dormant cysts and for avoiding resource intensive surveillance for benign cysts. We present a model that is capable of differentiating between three types of pancreatic cysts with an overall accuracy of 69.8%.

1. Introduction

Over the past few years, there have been many advances to classifying tumours with machine learning approaches. The current field of study that looks at extracting quantitative features from medical images, termed *radiomics*, is moving to using more complicated network architectures with medical imaging.

However, one of the problems with radiomics is that baseline results aren't that accurate. Simultaneously, the lack of readily-interpretable explanations behind deep learning models hinders their adoption by medical professionals, who are reluctant to depend on black box models. Additionally, many of the techniques that are currently employed in radiomics use 2D slices of the image which prevents the full usage of all information from 3D medical images.

Particularly for the problem of classifying cancerous pancreatic tumors, there are several different pancreatic cystic lesions that are hard to differentiate from each other. If these tumors could be accurately classified then different treatment plans for the respective tumors could be laid out earlier and lead to more successful recovery and faster detection.

Specifically, we examine the problem of identifying pancreatic tumors based on X-ray computed tomography (CT) scans. Unlike conventional X-ray scans, a CT scanner rotates around a patient, taking a volumetric series of 2D slices of a target area. The increasing usage of CT scans have raised concerns on the radiation exposed to patients in the scanning process. This has led to the principle of "as low as reasonably achievable" (ALARA). However, the reduced radiation dosage unfortunately leads to noisier images. The volumetric and potentially noisy nature of CT scans present unique challenges when we use CT scans as inputs. Due to the ubiquity of CT scans in clinical applications, a secondary objective of our project was to examine and exploit properties of CT scans.

In this paper, we follow our two objectives to understand and explain why different network architectures perform as they do. Our contributions are as follows:

- Usage of Inception Net Architecture in pre-training
- Specific techniques to deal with CT inputs (volumetric, denoising)
- Working on a novel medical imaging problem with a novel dataset.

2. Related Work

In recent years, there has been an explosion of deep learning techniques applied to radiology applications. Due to the fact that radiology is a field heavily reliant on extracting information from often large collections of images, it is an ideal application of deep learning. We refer the reader to the excellent survey paper of Mazurowski1 et al [6] for a more comprehensive collection of applications.

Early works in classifying CT scans as in the work of [4] did not directly deal with images, but instead a large corpus of features from each image (e.g. mean, standard deviation, skewness, kurtosis, n'th central moments, etc) and use these as inputs to a neural network. Recently, Tajbakhsh et al were the first to show that a pretrained convolutional neural net (CNN) upon fine tuning perform as well as or outperforms a CNN trained from scratch on four distinct medical imaging applications that did not include CT scans [9]. A number of further works [10], [5], [11] have similarly utilized transfer learning techniques with varying levels of success, on applications such as prostate cancer identification, pulmonary tuberculosis identification and hip osteoarthritis diagnosis. However, none of these works seemingly leverage the 3-D nature of CT scans nor do they process high

noise, low dosage CT scans. Furthermore, as far as we are aware, the specific domain of pancreatic cancer classification and our dataset in particular, has not been explored before with a deep learning approach.

3. Dataset

We are using segmented CT scans of three different pancreatic tumor classes from the Memorial Sloan Kettering Cancer Center (MSKCC). The three classes are pancreatic ductal adenocarcinoma (PDAC), intraductal papillary mucosal neoplasms (IPMNs), and pancreatic neuroendocrine tumors (PNETs). There are no benign scans. The dataset consists of 103 patients with IPMN, 57 with PNET, and 260 with PDAC, for a total of 420 patients. This is one of the largest pancreatic imaging datasets available, but is still limited when compared to traditional computer vision datasets.

4. Approach

Our approach to this problem is applying deep learning in the form of convolutional neural networks for this classification task. The usage of neural networks for Radiomics has started fairly recently and many results aren't that good. We think this convolutional neural network approach will work better because of its more complicated architecture and ability to differentiate based on studies in other domains. While this approach is not novel because of recent concurrent research, it can still achieve state-of-the-art for some problems in this category, and this paper goes over different network architectures and explanations which is certainly novel in this emerging field.

4.1. Preprocessing

The CT scans in the data set are stored as 3D MetaImage (MHD) files. In order to apply transfer learning and be able to use image classification neural networks, we need to extract 2D images of the tumor. The data were split in a 70/15/15 ratio into training/validation/test sets at the patient level, so that all the slices from a given volumetric scan were all distributed into the same dataset.

4.2. Baseline model

We took the Inceptionv3 model that was pretrained on ImageNet and fine-tuned it on 2D slices of the tumors. The scans were preprocessed to only include voxels with Hounsfield units between -100 and 300, which was then scaled to 0-255. The segmented tumors were resized by the same constant to 224x224 pixels such that the largest segmented tumor occupied the entire region and smaller tumors stayed the same size proportionally. This was then resized to 299x299 pixels when passed into the Inception model. The entire model was trained for 20 epochs using stochastic gradient descent with momentum (learning rate=0.001 decaying by a factor of 2 every 15 epochs, momentum=0.9) with batch size 64 and the parameters with the best validation accuracy was saved. The last layer was replaced with a series of 3 linear layers that brought the number of output features to 500, 20, and then finally to 3 output classes.

We also implemented a two-stage model to attempt to improve on the baseline by first classifying PDAC vs. non-PDAC, then classifying the non-PDAC classes as either IPMN or PNET.

4.3. Volumetric Neural Network

An alternative to using 2D slices of the CT scan, is to extract 3D images of the tumor and train a volumetric neural network or 3d convolutional neural network. [7] uses a volumetric convolutional neural network to perform segmentation of MRI images. This type of network is also advantageous for classification because features along the depth of the CT scan can be learned. The whole 3D scan has more information than one slice.

To create the volumetric data, the segmentation masks are applied to each slice of the scan. The slices that contain the tumor are adjacent with depth and cropped and concatenated into one 3D image, which is then resized to (64,64,16). This creates a dataset of 426 3D images. It is important to note that this dataset has much fewer examples, since one patient can only produce one sample rather than many 2D slices.

4.4. RED-CNN

Another approach we tried is based on the RED-CNN architecture due to Chen et al [1]. The original objective of the RED-CNN paper to address the noisy nature of low-dosage CT scans by using an encoder - decoder pair to reconstruct the original images. We wanted to attempt using this architecture as a form of anomaly detection - to measure the reconstruction error and marking any image with abnormally high reconstruction error as an anomaly. In practice, this approach may be more robust when dealing with rare classes of tumors and could be robust when encountering previously unseen tumors. Our implementation mirrors that of Chen et al, we use the architecture specified in Chen et al's paper and use an Adam optimizer with learning rate of 0.001.

5. Experiments and Results

We ran experiments using different Network Architecture decisions by splitting the data into a training, validation, and testing set. After training the model with varying hyperparameters and augmentations of the CT images, we calculated testing accuracy and present the results below.

		Predicted		
		IPMN	PNET	PDAC
Actual	IPMN	16	0	0
	PNET	0	9	0
	PDAC	9	10	18

Table 1. Confusion matrix of one-stage network on test set

		Predicted	
		PDAC	Non-PDAC
Actual	PDAC	21	18
	Non-PDAC	1	23

Table 2. Confusion matrix of two-stage network on classifying PDAC vs. non-PDAC

		Predicted	
		IPMN	PNET
Actual	IPMN	16	0
	PNET	0	9

Table 3. Confusion matrix of two-stage network on classifying IPMN vs. PNET

5.1. Baseline Model

Overall, the baseline has good performance: it has 100% sensitivity on PDAC and IPMN, but performance drops when attempting to distinguish PDAC from the other classes. Sensitivity for PDAC on the test set is 45%. This brings overall accuracy on the test set to 69.3%.

The two-class model is slightly better: it maintains 100% test accuracy on the second stage but still does poorly on the first stage. On the two-class PDAC vs. non-PDAC problem, we have a sensitivity of 53.8% and a specificity of 95.8% with PDAC as the positive class.

We also wanted to examine the quality of the features extracted from our network. We did this by applying t-SNE on the output of a fine-tuned network using the PNASNet-5 large architecture. Visually, it appears that IPMN and PNET live in separate areas of feature space from each other, but overlap heavily with PDAC. This shows the difficulty in the classification.

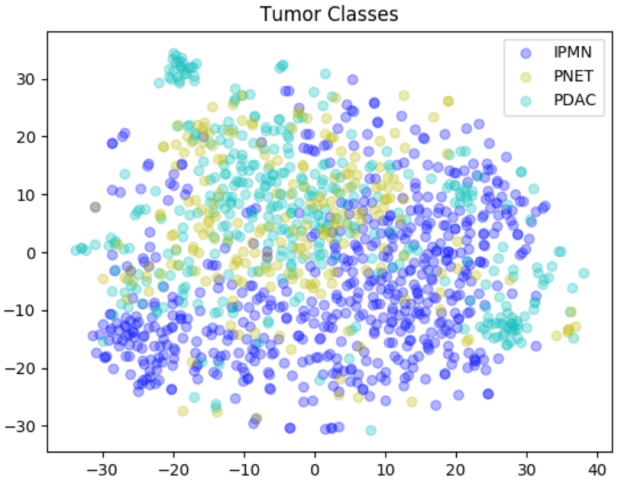

Figure 1. t-SNE Visualization of Embeddings

5.2. Volumetric Neural Network

We implemented a network with three 3D convolutional layers and two fully connected layers for classification. The dataset was divided into training and test sets with a 67-33 split. The model was trained using Adam[3] optimizer with a learning rate of 0.001 for 20 epochs. This network was able to achieve a classification accuracy of 91% for PDAC and 85% for IPMN. The network is not able to classify any of the PNET samples. This approach is very limited by the training data. Because each scan only results in one 3D tumor image, a dataset with large scans would be needed to more effectively use volumetric neural networks. However, this is a promising direction because the network is far smaller and more computationally efficient.

5.3. RED-CNN

Due to the lack of healthy patients in our data set, we trained our RED-CNN on our PDAC as it was the class with the largest number of samples. After training the RED-CNN for 50 epochs, we attempted to see if reconstruction error on PDAC vs non-PDAC classes were significantly different. Unfortunately, there was no measurable difference - the reconstruction errors were roughly the same even across varying code layer sizes.

5.4. Grad-CAM

We used Gradient weighted Class Activation Maps (GRAD-CAM) [8] to look at the gradient weighted channels to help understand what the Convolutional Neural Networks are perceiving. The neural networks have no problem detecting the presence of tumors but it is hard to interpret what weights make the neural network decide on the class of the tumors since to the untrained and unprofessional human eye, many of the tumors look the same and are small in size. However, Grad-CAM does show us that different tumor classes have different activations for the most part.

		Predicted		
		IPMN	PNET	PDAC
Actual	IPMN	28	0	5
	PNET	1	0	17
	PDAC	7	0	78

Table 4. Confusion matrix of 3D-CNN classifier

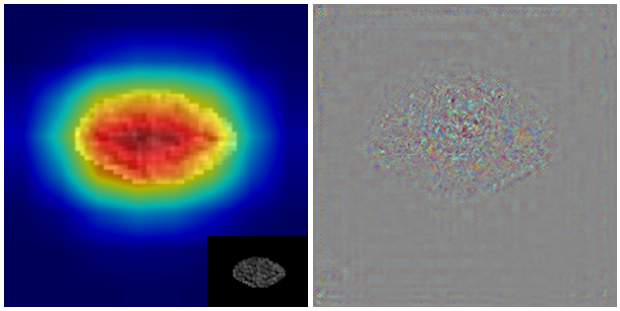

Figure 2. Grad-CAM of example PDAC tumor with scaled original image superimposed on the bottom

6. Discussion

One limitation of our work is that our analysis is retrospective, using already collected data. Furthermore, there may be additional selection bias from the fact that these were collected from a cancer center: benign cysts are not present in the dataset, and the precancerous PNET and IPMN tumors present were all suspicious enough to have been resected and confirmed by pathology. Our model is exceptional at correctly identifying IPMN and PNET when present, however, which leads to a positive predictive value of 100% for PDAC. This is biased in the wrong direction, though–one would prefer more false positives of PDAC rather than false negatives, as PDAC is actively malignant and the cost is more severe if misclassified.

7. Conclusion

Given the results from our experiments, there is noticeable benefit to using convolutional neural networks for the task of pancreatic tumor classification especially for classifying tumors as IPMN or PNET. However, there is still much more work needed to be done in order to correctly classifying all classes of pancreatic tumors accurately enough to be useful in a realistic setting.

7.1. Future Direction

A potential future direction of research in classifying pancreatic tumors is low shot learning, especially when dealing with rare tumors. This was actually what we originally hoped our project would focus on. Unfortunately, while attempting to implement a low shot learning approach, we discovered that there is a lack of comprehensive data on cancer patients to use as side information (due to patient privacy laws) and it is also hard to find data with non-cancerous patients.

8. Acknowledgments

We greatly thank the Amber Simpson lab at Memorial Sloan-Kettering Cancer Center for advice and for providing the dataset.

References

[1] Hu Chen et al. “Low-Dose CT with a Residual Encoder-Decoder Convolutional Neural Network (RED-CNN)”. In: *arxiv preprint* ().

[2] Linda Chu, Seyoun Park, and Elliott Fishman. “AB032. P002. Radiomics based classification of pancreatic cystic neoplasms”. In: *Annals of Pancreatic Cancer* 1.4 (2018). URL: http://apc.amegroups.com/article/view/4210.

[3] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014).

[4] Jinsa Kuruvilla and K. Gunavathi. “Lung cancer classification using neural networks for CT images”. In: *computer methods and programs in biomedicine* ().

[5] Paras Lakhani and Baskaran Sundaram. “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks.” In: *Radiology* ().

[6] Maciej A. Mazurowski et al. “Deep learning in radiology: an overview of the concepts and a survey of the state of the art”. In: *arXiv preprint* ().

[7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *CoRR* abs/1606.04797 (2016).

[8] Ramprasaath R. Selvaraju et al. “Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization”. In: *CoRR* abs/1610.02391 (2016). arXiv: 1610.02391. URL: http://arxiv.org/abs/1610.02391.

[9] Nima Tajbakhsh et al. “Convolutional neural networks for medical image analysis: Full training or fine tuning?” In: *IEEE transactions on medical imaging* ().

[10] Xinggang Wang et al. “Searching for prostate cancer by fully automated magnetic resonance imaging classification: Deep learning versus non-deep learning.” In: *Scientific Reports,* ().

[11] Yanping Xue et al. “A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis”. In: *PLOS ONE* ().