國立臺灣大學理學院應用數學科學研究所

碩士論文

Institute of Applied Mathematical Science

College of Science

National Taiwan University

Master Thesis

遷移學習應用於
二維胰臟影像小區塊方式之腫瘤辨識

Applying Transfer Learning on 2D Patch-Based Healthy Pancreas and Pancreatic Tumor Classification

楊宛芸

Wanyun Yang

指導教授：王偉仲 教授

Advisor: Weichung Wang, Ph.D.

中華民國 109 年 4 月 (24日)

Acknowledgement

I would first like to thank my thesis advisor, professor Weichung Wang. His lab provides rich learning resources and precious opportunities. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever I have difficulties in research.

I would also like to thank the doctors at National Taiwan University Hospital who were involved in the pancreatic project: Dr. Wei-Chih Liao and Dr. Poting Chen. Without their medical professionalism and advice on research method, the research could not have been successfully conducted.

I'm glad to thank a excellent research assistant Tinghui Wu of the Academia Sinica as the second reader of this thesis, and I am gratefully indebted to her for her very valuable comments on this thesis. She also kindly helped me for both programming and experiment setting.

Finally, I must express my very profound gratitude to my parents and to my friends Chunshuo Chen and YuYuan Yuan. At the end of the second semester of master degree, I was stranded at a bottleneck of my research. They provided me with practical support and encouragement throughout my dilemma and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

((Structured summary of study design, methods, results, and conclusions))

Pancreatic cancer is the second most frequent cancer of the digestive system, and it's an excellent choice to apply artificial intelligence to help doctors with tumor detection.

I'm in Dr. Weichung Wang's research team and work on this detection problem with other master students. But all medical images, including pancreas CT, are precious and hard to get. With the help of doctors in National Taiwan University Hospital (NTUH), we get enough data to train an accurate convolution neural network. In the future we need to apply our model to different dataset, but the performance of detection model decreases when testing on other datasets. Our goal is to enhance the performance of model on other dataset.

In this paper, we seek to answer the following core question in the context of medical image analysis: How much fine-tuning improves the performance on external dataset? How much data we need to get a certain performance? Our experiments consistently demonstrated that, as we get more and more external data, we can get a better prediction result on external data. We can at most enhance the area under the receiver operating characteristic curve (AUC) up to 0.04. This paper can help to evaluate how much external data we need to get a satisfied result and help our team to promote our neural network result worldwide.

Contents

1 Introduction **1**

1.1 Clinical background 1

1.2 Study objectives 2

2 Methods **3**

2.1 Study Design 3

2.1.1 Prospective or retrospective study 3

2.1.2 Study goal* 4

2.2 Data 4

2.2.1 Data Source 4

2.2.2 Eligibility criteria* 5

2.2.3 Data Preprocessing 5

2.2.4 Selection of data subsets(ignore) 6

2.2.5 Definitions of data elements(ignore) 6

2.2.6 De-identification Methods* 6

2.2.7 Missing Data* 6

2.3 Ground Truth* 6

2.3.1 Reference Standard 6

2.3.2 Rationale for choosing the reference standard 6

2.3.3 Source of Ground Truth Annotations 6

2.3.4 Annotation tools 7

2.3.5 Measurement of inter and intrarater variability 7
2.4 Data Partitions* 7
2.4.1 Sample Size and How It Was Determined 7
2.4.2 Assign to partitions 7
2.5 Model 7
2.5.1 Detailed description of model 7
2.5.2 Software libraries, frameworks, and packages* 8
2.5.3 Initialization of model parameters* 8
2.6 Training* 9
2.6.1 Training approach 9
2.6.2 Method of selecting the final model 9
2.6.3 Ensembling techniques, if applicable(ignore) 9
2.7 Evaluation 9
2.7.1 Model performance 9
2.7.2 Statistical measures of significance and uncertainty* 11
2.7.3 Robustness or sensitivity analysis* 11
2.7.4 Methods for explainability or interpretability* 11
2.7.5 Validation or testing on external data(ignore) 12
2.7.6 Area under the Receiver Operating Characteristic Curve (AUC)? 12
2.7.7 Transfer Learning and Fine-Tuning? 12
2.7.8 Cross Validation? 12

3 Results **14**
3.1 Data* 14
3.1.1 Flow of participants or cases 14
3.1.2 Demographic and clinical characteristics of cases 14
3.2 Model performance 15
3.2.1 Performance metrics for optimal model 15
3.2.2 Estimates of diagnostic accuracy and their precision* 17

3.2.3 Failure analysis* 18

4 Discussion 19

4.1 Limitations 19

4.2 Implications for practice* 19

4.3 Future Work? 19

5 Other* 20

5.1 Information 20

5.1.1 Registration number and name of registry 20

5.1.2 Where the full study protocol can be accessed 20

5.1.3 Sources of funding and other support; role of funders 20

Chapter 1

Introduction

1.1 Clinical background

((Scientific and clinical background, including the intended use and clinical role of the AI approach))
Pancreatic cancer is the second most frequent cancer of digestive system which caused 45,750 deaths in the United States in 2019[SMJ19]. The research team of my thesis advisor Dr. Wang started working on pancreatic cancer since 2018 and had developed 2D patch-based tumor classification model based on patient data from National Taiwan University Hospital(NTUH). But when we tested our model using data from other data sources, the performance of the classification model is worse than the previous results. (In this thesis, we mainly use the area under receiver operating characteristic curve (AUC) to evaluate the performance.) Our goal is to improve the performance of model working on external data using fine-tuning technique so that we can apply our model to patient data from different datasets and different countries in the future.

1.2 Study objectives

((Study objectives and hypotheses))

Consider a source dataset and an external dataset. Source dataset has sufficient data to train a model, and external dataset doesn't have sufficient data to train a workable model. The purpose of the thesis is

- Use source data and target data to build a model that has high performance on target test set.
- Evaluate how many patients are needed to improve the performance.

Chapter 2

Methods

2.1 Study Design

2.1.1 Prospective or retrospective study

((Prospective or retrospective study))

With the revival of neural networks owing to the development of parallel computing technique, the medical imaging field has witnessed a new generation of computer-aided systems that show incredible performance. A Survey on Deep Learning in Medical Image Analysis [LKB$^+$17]surveyed the use of deep learning techniques such as image classification, object detection, segmentation, registration.

One important property of CNNs is the "transferability". Due to the efforts of the doctors in NTUH, we have sufficiently large pancreatic dataset. We want to use this precious data to help train models for other pancreatic dataset. It deeply relies on the "transferability" of CNN, in other words, transfer-learning.

But there are multiple transfer-learning methods, why we choose fine-tuning? According to this thesis Learning without Forgetting [LH17], the authors compare four methods, including joint training, feature extraction, learning without forgetting and fine-tuning. The fine-tuning method performs well in external data, and also has excellent testing efficiency. Another thesis in 2017 applies fine-tuning on ultrasound images and has an obvious enhance on model performance. [CWB+17]

Since 2018, our team continues collecting new external data and labeling the images. If we can select patient data that provides information most, the model performance increases more for certain amounts of labeled data. [ZSZ+17] provides AIFT method to do data selection.

The thesis applies fine-tuning model on pancreatic images, also data selection will be applied.

2.1.2 Study goal*

((Study goal, such as model creation, exploratory study, feasibility study, noninferiority trial))

2.2 Data

2.2.1 Data Source

((Data sources))

This thesis applies three datasets. All of them are pancreatic CT images.

- **National Taiwan University Hospital (ntuh, source data)**

  including both healthy(400) and tumor(400) CT scans.

- **Medical Segmentation Decathlon[SAB+19] (msd, target data)**

  including only tumor(281) CT scans.

- **The Cancer Imaging Archive (tcia, target data)**

  including only healthy(82) CT scans.

2.2.2 Eligibility criteria*

((Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (eg, symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)))

2.2.3 Data Preprocessing

((Data preprocessing steps))

All CT images that contained the pancreas or PC from an individual subject were manually labeled for further model training/validation and testing using an open source software (3D Slicer version 4.8.1). Since the pancreas bordered multiple organs/structures and PCs often had an indistinct border with the surrounding tissue, there are inter-observer differences regarding the exact extent of the pancreas and the cancer tumor. Therefore, the labeled pancreas and tumor on the images were checked by the radiologists before further processing and analysis steps. The window width and window level were fixed as 250 Hounsfield unit (HU) and 75 HU, respectively. The images were normalized to [0, 1] by linear interpolation, and the portions that were neither pancreas nor tumor were excluded from further analysis. The images were then cropped into square sub-regions (i.e. 50 X 50 patches) using the moving window method on the axial (x-y) plane, starting from the top-left corner and ended at the bottom-right corner. Moving distance was set as half of the patch dimension to generate overlapping patches in order to increase the variation and size of training data. The patches which contained PC were labeled as cancerous, whereas patches that contained only non-cancerous pancreatic parenchyma were labeled as non-cancerous.

2.2.4 Selection of data subsets(ignore)

((Selection of data subsets, if applicable(ignore)))

2.2.5 Definitions of data elements(ignore)

((Definitions of data elements, with references to common data elements(ignore)))

2.2.6 De-identification Methods*

((De-identification methods))

2.2.7 Missing Data*

((How missing data were handled(ignore)))

2.3 Ground Truth*

2.3.1 Reference Standard

((Definition of ground truth reference standard, in sufficient detail to allow replication))

2.3.2 Rationale for choosing the reference standard

((Rationale for choosing the reference standard (if alternatives exist)))

2.3.3 Source of Ground Truth Annotations

((Source of ground truth annotations; qualifications and preparation of annotators))

2.3.4 Annotation tools

((Annotation tools))

2.3.5 Measurement of inter and intrarater variability

((Measurement of inter and intrarater variability; methods to mitigate variability and/or resolve discrepancies))

2.4 Data Partitions*

2.4.1 Sample Size and How It Was Determined

((Intended sample size and how it was determined))

2.4.2 Assign to partitions

((How data were assigned to partitions; specify proportions))

2.5 Model

2.5.1 Detailed description of model

((Detailed description of model, including inputs, outputs, all intermediate layers and connections))

The CNN model was modified from VGG network12, a neural network widely used in image classification. Weighted binary cross-entropy13 was used as the loss function to solve the imbalance problem between the number of cancerous and non-cancerous patches. Table 1 provides the details of the CNN model, including the layer structures, kernel sizes, channels, and output sizes of the network. Two callbacks monitoring on validation loss were used during the training process to optimize model performance. All codes were written in Python (version 3.6.8) using Keras (version 2.2.4)14 and Tensorflow (version 1.7.0)15 libraries.

Table 2.1: CNN model Structure

Layer (type)	Output Shape	Parameter
conv2d 1 (Conv2D)	(None, 50, 50, 16)	416
conv2d 2 (Conv2D)	(None, 50, 50, 32)	12832
max pooling2d 1 (MaxPooling2)	(None, 50, 50, 32)	0
conv2d 3 (Conv2D)	(None, 25, 25, 64)	18496
conv2d 4 (Conv2D)	(None, 25, 25, 64)	36928
max pooling2d 2 (MaxPooling2)	(None, 12, 12, 64)	0
conv2d 5 (Conv2D)	(None, 12, 12, 128)	73856
conv2d 6 (Conv2D)	(None, 12, 12, 128)	147584
max pooling2d 3 (MaxPooling2)	(None, 6, 6, 128)	0
flatten 1 (Flatten)	(None, 4608)	0
dense 1 (Dense)	(None, 32)	147488
dropout 1 (Dropout)	(None, 32)	0
dense 2 (Dense)	(None, 32)	1056
dense 3 (Dense)	(None, 1)	33

2.5.2 Software libraries, frameworks, and packages*

((Software libraries, frameworks, and packages))

2.5.3 Initialization of model parameters*

((Initialization of model parameters (eg, randomization, transfer learning)))

2.6 Training*

2.6.1 Training approach

((Details of training approach, including data augmentation, hyperparameters, number of models trained))

2.6.2 Method of selecting the final model

((Method of selecting the final model))

2.6.3 Ensembling techniques, if applicable(ignore)

((Ensembling techniques, if applicable(ignore)))

2.7 Evaluation

This chapter consists of three experiments: basic validation, mix data training and fine tuning. To better observing the result, this thesis use cross validation to distribute training/validation/testing set.

2.7.1 Model performance

((Metrics of model performance))

Basic Validation

In this part, we train model only using target data. We want to find how many patients in target data is needed to build a sufficiently nice model. That means, the

AUC of model is larger than 0.8. The table below is the diagram of the number of training/validation/testing set of this experiment. We increase the number of target data to observe the performance of model trained by target data (82, 164, 246, 326 target data).

Table 2.2: training/validation/testing set (A) (Basic Validation, 10 folder)

	source H	source T	target H (tcia)	target T (msd)
train	0	0	-	-
validation	0	0	-	-
source test	40	40	0	0
target test	0	0	8	28

Table 2.3: training/validation/testing set (B) (10 folder)

amount of data	tar H (train)	tar T (train)	tar H (val)	tar T (val)
82	17	57	2	6
164	33	114	4	13
246	50	171	6	19
326	66	228	7	25

Mix Source and Target Data

In this part, we train model using both source data and target data. We want to find how the number of patients in target data influences the model performance. The table below is the diagram of the number of training/validation/testing set of this experiment. We increase the number of target data to observe the performance of joint training (82, 164, 246, 326 target data).

Table 2.4: training/validation/testing set (Mix Source and Target Data)

	source H	source T	target H (tcia)	target T (msd)
train	324	324	-	-
validation	36	36	-	-
source test	40	40	0	0
target test	0	0	8	28

Transfer Learning

In this part, first we train model only using source data. Than I use the weight of the previous model as the initial value of training. I train the previous model again using target. The table below is the diagram of the number of training/ validation/ testing set of this experiment.

Table 2.5: training/validation/testing set (A) (Basic Validation, 10 folder)

	source H	source T	target H (tcia)	target T (msd)
train	0	0	-	-
validation	0	0	-	-
source test	40	40	0	0
target test	0	0	8	28

Table 2.6: training/validation/testing set (B) (10 folder)

amount of data	tar H (train)	tar T (train)	tar H (val)	tar T (val)
82	17	57	2	6
164	33	114	4	13
246	50	171	6	19
326	66	228	7	25

2.7.2 Statistical measures of significance and uncertainty*

((Statistical measures of significance and uncertainty (eg, confidence intervals)))

2.7.3 Robustness or sensitivity analysis*

((Robustness or sensitivity analysis))

2.7.4 Methods for explainability or interpretability*

((Methods for explainability or interpretability (eg, saliency maps) and how they were validated))

2.7.5 Validation or testing on external data(ignore)

((Validation or testing on external data(ignore)))

2.7.6 Area under the Receiver Operating Characteristic Curve (AUC)?

A receiver operating characteristic curve, or ROC curve, is a graphical plot that shows the diagnostic ability of a binary classifier system as its predict threshold is varied. The area under the receiver operating characteristic curve is a common criteria to evaluate model performance, not depending on patient threshold. We use AUC in our this thesis.

2.7.7 Transfer Learning and Fine-Tuning?

In real-world applications, the assumption that the training and future data must be in the same feature space and have the same distribution may not hold. In such cases, knowledge transfer, if done successfully, would highly improve the performance of learning by avoiding much expensive data-labeling efforts. In medical image research, patient data and label is precious and hard to get. Transfer learning has emerged as a new learning framework to solve this problem. Since the target data is labeled, the thesis use fine-tuning to enhance the prediction performance.[PY09]

2.7.8 Cross Validation?

Cross-validation is any of various similar model validation techniques for assessing how the results of a model will generalize to an independent data set.One wants to estimate how accurately a predictive model will perform in practice. The goal of cross-validation is to test the model's ability to predict new data that was excluded from training data, in order to deal with problems like overfitting or selection bias.

Take 5-folder cross validation as an example. There are 400 healthy ntuh data. First we split it into 5 groups (80 data for each group) and take them as test set. 90 percent of the remained data is the training data while 10 percent is the validation data. (Figure 3.1)

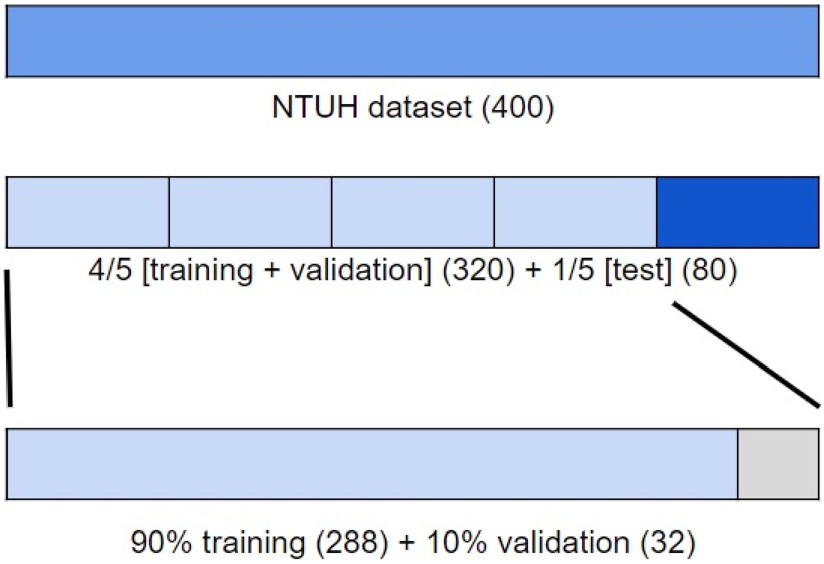
NTUH dataset (400)
4/5 [training + validation] (320) + 1/5 [test] (80)
90% training (288) + 10% validation (32)

Figure 2.1: training/validation/testing set distribution

Chapter 3

Results

In order to validate the performance, I use 10-folder cross validation to do all the experiments .

3.1 Data*

3.1.1 Flow of participants or cases

((Flow of participants or cases, using a diagram to indicate inclusion and exclusion))

3.1.2 Demographic and clinical characteristics of cases

((Demographic and clinical characteristics of cases in each partition Model performance))

3.2 Model performance

3.2.1 Performance metrics for optimal model

((Performance metrics for optimal model(s) on all data partitions))

Basic Validation

In this experiment, the AUC of experiments increases as the number of training data grows. But the AUC is not high enough for practical use.

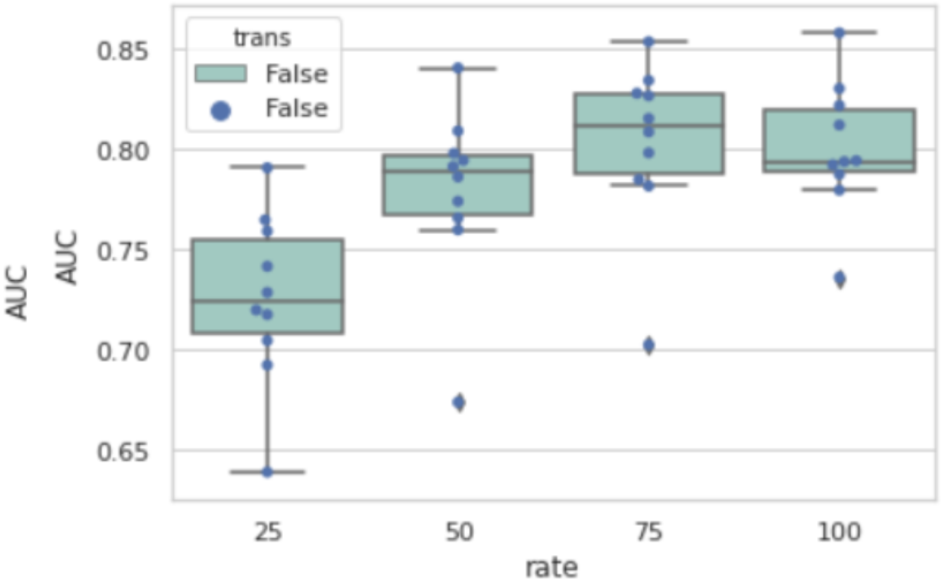

Figure 3.1: AUC for Basic Validation

Mix Source Data and Target Data

In this experiment, with the help of source data, the AUC of experiments increases as the number of target data grows. The performance increases most when the number of target data is small.

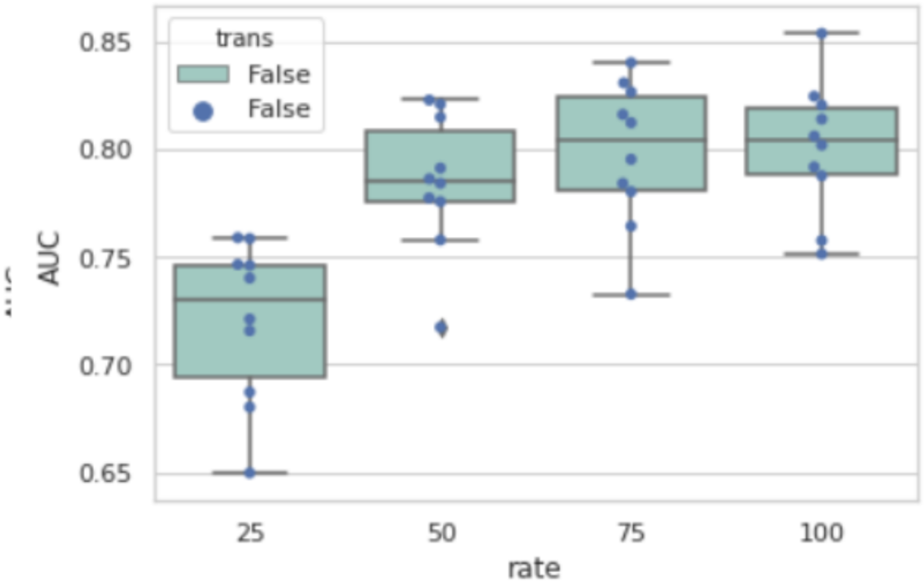

Figure 3.2: AUC for Mix Source Data and Target Data

Fine Tuning

In this experiment, I try two methods: fine-tuning with no layer fix and with 3 layers fixed. The AUC of experiments increases as the number of target data grows. The performance of 3 layer-fixed experiment is slightly better than 0 layer-fixed experiment.

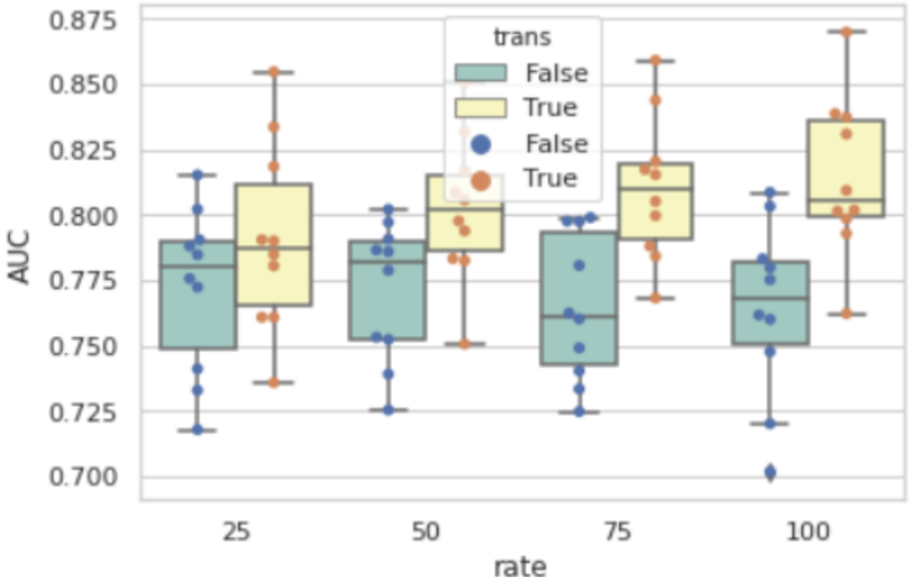

Figure 3.3: AUC for Fine Tuning (fix 0 layer)

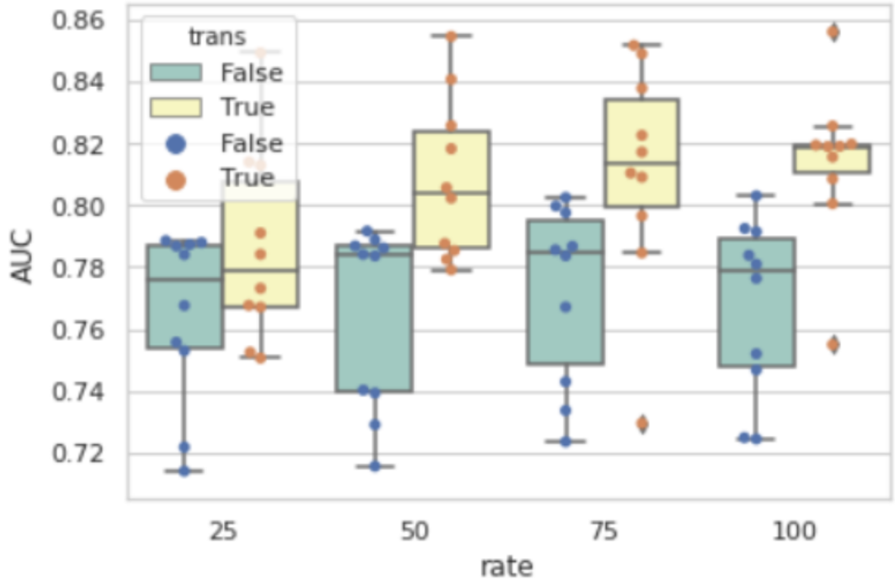

Figure 3.4: AUC for Fine Tuning (fix 3 layer)

3.2.2 Estimates of diagnostic accuracy and their precision*

((Estimates of diagnostic accuracy and their precision (such as 95 percent confidence intervals)))

3.2.3 Failure analysis*

((Failure analysis of incorrectly classified cases))

Chapter 4

Discussion

4.1 Limitations

((Study limitations, including potential bias, statistical uncertainty, and generalizability))
Since the target data is composed of two dataset: MSD dataset and TCIA dataset, the distribution of two datasets might be different, so the model performance may decrease because of the reason. Also, since the number of data is not large enough, the outliers shows up in cross validation experiments.

4.2 Implications for practice*

((Implications for practice, including the intended use and/or clinical role))

4.3 Future Work?

We can try different fine-tuning details of AIFT method. It's possible to increase the performance on target data.

Chapter 5

Other*

5.1 Information

5.1.1 Registration number and name of registry

5.1.2 Where the full study protocol can be accessed

5.1.3 Sources of funding and other support; role of funders

Bibliography

[CWB$^+$17] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486, 2017.

[LH17] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[LKB$^+$17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[SAB$^+$19] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[SMJ19] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[ZSZ+17] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.