Prostate cancer diagnosis and gleason grading using histological images

By ANIMESH PAL

WORD COUNT	4205	TIME SUBMITTED	21-JUL-2022 11:50AM
		PAPER ID	88647034

Prostate cancer diagnosis and gleason grading using histological images

A thesis submitted in partial fulfillment of the requirements for the award of the degree of

M.Tech

in

Computer Sciecne

By

Animesh Pal (Roll No. CS2008)

Project Guide : PROF. Pradipta Maji

Machine Intelligence Unit
Indian Statistical Institute
Kolkata, West Bengal – 700108

JULY 2022

CERTIFICATE

This is to certify that the dissertation entitled **Prostate cancer diagnosis and gleason grading using histological images** submitted by

Animesh Pal (Roll No. CS2008)

to **Indian Statistical Institute, KOLKATA**, in partial fulfillment for the award of the degree of **Master of Technology** in **Computer Sciecne** is a bonafide record of work carried out by him under my supervision and guidance.

PROF. Pradipta Maji
Supervisor
Machine Intelligence Unit

Date ________________________

ACKNOWLEDGEMENTS

I would like to express our deepest gratitude to the following people for guiding us through this course and without whom this project and the results achieved from it would not have reached completion.

PROF. Pradipta Maji, Professor, Machine Intelligence Unit,Indian Statistical Institute, Kolkata, for helping us and guiding us in the course of this project. Without his guidance, we would not have been able to successfully complete this project. His patience and genial attitude is and always will be a source of inspiration to us.

We are also thankful to the faculty and staff members of the Department of Computer Sciecne, our individual parents and our friends for their constant support and help.

ABSTRACT

We present a study of tissue images for prostate cancer diagnosis and Gleason grading of the histological images of the prostate. In Prostate cancer diagnosis Gleason grading is most powerful predictor for patients since 1960s. Histological Images were captured from representative areas of Hematoxylin and Eosin(H&E) stained tissue retrieved from TMA(Tissue Micro Array) cores. Previously it is used to take the image features containing the color, texture and morphometric cues at the histological object level to classify the images. Nowadays after the invention of Deep learning Feature extraction and classification both happens in a single Deep Neural Network. In this study we used MobileNet architecture which is a Deep CNN to get an accuracy around 58 percent. Total TMA images used on 641 patients for training and 245 TMA images is taken as testing which is annotated by two expert pathologists to check with the model. We can see that Gleason score (The most two predominant gleason grade present in the images) assignments achieved approximately same Cohen's quadratic kappa score between the two pathologists.

Keywords : Gleason Grading, Prostate Cancer, Deep Learning

TABLE OF CONTENTS

Title **Page No.**

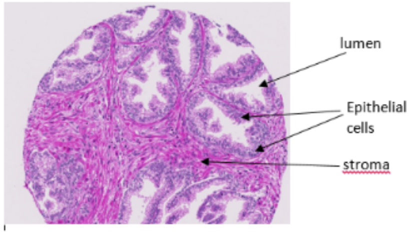
LIST OF FIGURES

CHAPTER 1

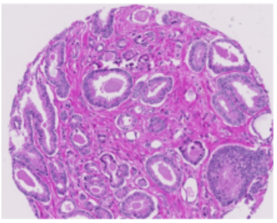
INTRODUCTION

1.1 Background

Prostate cancer is the most prevalent form of cancer and the second leading site of cancer among males in Indian cities like Kolkata, Delhi, Pune and Thiruvananthapuram, third leading site of cancer in cities like Bangalore and Mumbai[1]. There are many methods existed before to detect prostate cancer. One of the most reliable method is the examination of histlogical specimens under the microscope by an expert pathologist. In the H&E stained specimen we can see the glandular architecture of prostate tissue[Fig (a)] which is surrounded by fibromascular tissues called stroma. Stroma holds the gland units together and these gland units consists of rows of epithelial cells around lumen. When prostate cancer happens, this epithelial cells replicate in an uncontrollable manner. The lumens have filled with epithelial cells and the stroma has virtually disappeared[Fig 1.2].



(a) Grade-1



(b) Grade-3

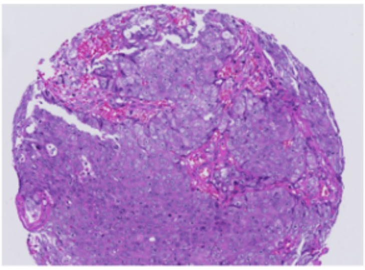


Figure 1.2: Grade-5

Aggressiveness of cancer is quantified through histological grading. This grading helps to identify the extent of the disease. It correlates well with survival analysis of the patient. Also it helps to determine the appropriate treatment options.[2]

One of the most popular method for histological grading is the Gleason grading system which is first invented by Donald Gleason in 1966[3][4]. In this system, the tissue is divided into 5 grades(from grade-1 to grade-5). Grade-1 is well-differentiated meaning that it behaves like a normal tissue where as grade-5 means poorly differentiated meaning that it corresponds to high replication cancerous regions. The patient with grade-5 has very low chance of survival.

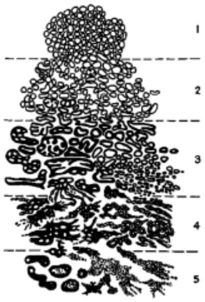


Figure 1.3: Gleason Grading Diagram

The gleason score which is the sum of most and second most predominant Gleason grades represents the heterogeneity of prostate cancer[5][6]. Gleason score(3+4) meaning that patient has most predominant grade-3 gleason grade and second most grade is grade-4. The Gleason score is ranging between 2 to 10.

But Gleason score annotation of TMA images is highly depended on the evaluation by an expert pathologist. This expert level annotation is highly time consuming and costly. So we have to have some model which predict expert level stratification of Gleason score. For prostate cancer detection Gleason score 7 is highly ambiguous since it is either (3+4) or (4+3)[7].

1.2 Previous work

In cancer diagnosis and Gleason grading is heavily subjective due to inter-pathologist reproducibility and it is related on human interpretation. Previously there are many feature extraction techniques have been developed on a small dataset. Diamond et al.[8] used morphometric and texture features to identify the possible cancerous regions given by large number of epithelial cells which replicating uncontrollably. They used Haralick texture feature to classify stroma and cancerous regions to achieve an accuracy of 79.3% on their dataset which is relatively biased.

There are many computer assisted Gleason Grading methods which extracts statistical properties and structural features from distribution of nuclei cells. Strotza et al.[9] used a hybrid Neural Network which is a Gaussian statistical classifier to classify histological samples with an accuracy of 77% on a set of 130 independent test images.

There are also some non-quantitative methods which uses features from spanning trees which connects nuclei cells across the tumor images to represent tissue of different grades.[10]

Tabesh et al.[11] used 268 color images for the diagnosis and Gleason grading using MAGIC feature extraction method to aggregate color texture and morphometric details at the global and histological object levels for classification. These features then channel through some supervised learning framework such as Gaussian, k-nearest neighbour, and support vector machine and together with sequential forward search selection algorithm. They achieved accuracy around 81% on Gleason grading classification of high grade and low grade tumor.

As opposed to overcome conventional feature extraction and classification, in recent years, deep learning[12] based approach are suitable alternative to feature based techniques. First Deep learning based approach is studied by Kallen et al[13]. They used Deep CNN architecture pre-trained on a large set of photographic images to detect and classify cancerous tissue. They used random forest classifier and a SVM classifier at the output layer with a dataset consisting of 213 images continging single class(homogeneous tissue) only. Accuracy of their result was around 80%.

Recently Zhou et al.[14] proposed a deep learning approach to test on 368 whole slide images from the TCGA(The Cancer Genoome Atlas) dataset with an accuracy of 75% in differentiating heterogeneous Gleason pattern of Gleason score 7(which is either (4+3) or (3+4)).

CHAPTER 2

METHODOLOGY

2.1 Introduction

Previously the system block diagram for classification problem consist of three stages- 1.Preprocessing, 2. Feature Extraction and then 3. Classification. But nowadays after the invention of Deep Learning Feature Extraction and classification both happens inside a Deep multiple layers of Neural network shown in Fig .

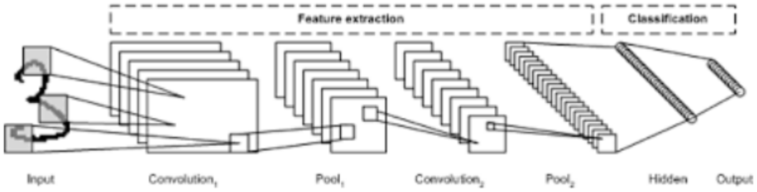


Figure 2.1: Deep Convolutional Neural Network architecture

2.2 DATA PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning or deep learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning or deep learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task. A real-world data generally contains noises and maybe in an unusable format which cannot be directly used for deep learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a deep learning model which also increases the accuracy and efficiency of a deep learning model. It involves below steps:

1. Getting the dataset
2. Importing libraries
3. Importing and Loading the datasets into data frame
4. Finding Missing Data

5. Our image datasets are huge in size (3100x3100) in pixels so fast we have to reduce the image so that it inputs in a DEEP CNN model.

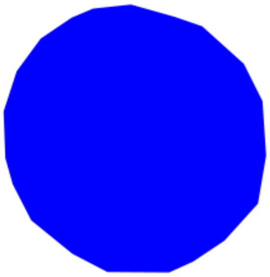
6. Appropriately find the labelling of the data

7. Splitting dataset into training, validation and testing set.

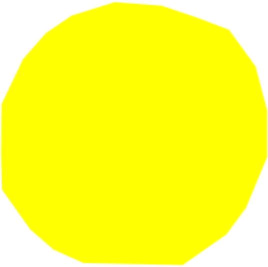
2.2.1 Datasets and labels

Our datasets contains in total 886 H&E stained histological images each of size of 3100x3100 pixels. These are the TMA(Tissue Microarray Images). Of these images 641 images are taken as training images and 245 images are taken as testing images. Images have different names TMA 111, 199, 204 are taken as training set, TMA 76 as validation set and TMA 80 as test set. In the Harvard Database where we find the data it is not mentioning in a csv format that these images have these gleason grades. The labels used in this datasets have 4 different colors with masking; green: Benign region, blue: Gleason 4 region, red: Gleason 5 region. Below are some example pictures. Two expert pathologist have also masked TMA 80 images in a masked color labelling way.

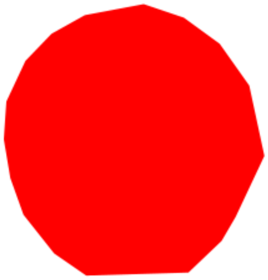
(a) Benign



(b) Gleason 3

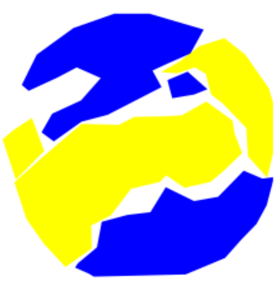


(a) Gleason 4

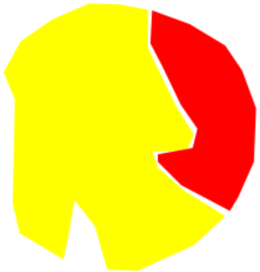


(b) Gleason 5

Also there are heterogeneous gleason pattern present in the data. Below are the masked images with color label as before.



(a) Gleason score (4+3)=7



(b) Gleason score (4+5)=9

2.2.2 Creating Patches

The first step in pre-processing is to create images patches to find the unique grading of the patches. These patches are used in training and validation. Step by step algorithm to create patches are bellow:

1. Here we have taken a patch size of 750. From the 3100x3100 images and masked labels we want to generate overlapping patches.

2. From the masked color patches we get patch level by using appropriate color palette as (0,255,0)-green as benign, (0,0,255) Blue as Gleason 3,(255,255,0) yellow as Gleason 4, (255,0,0)-red as Gleason 5 and (255,255,255)-white as ignore class.

3. Get patch level by creating a window size as patch size/3. If the central grades found in the window is 1 i.e. unique grade can assigned by using the color palette. We take only those patches with unique grade. This process also have been taken for test images as well. For the test case we took the patches with levels when for both pathologist annotated mask in that window have unique grade.

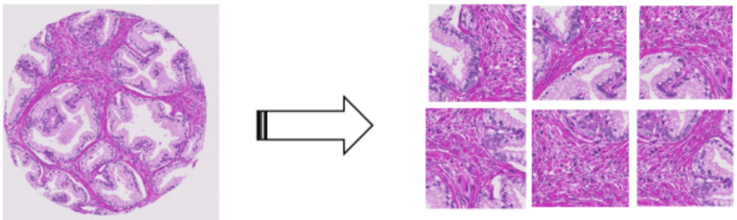


Fig: Original Image

Image Patches

4. Read the whole mask and count the grade at each patch and get the max and second max grades taken into a info directory. In this way we find each image labelled as 2 predominant grades. For the test data there have been two predominant grades with two pathologist label.

5. During this patch creation we ignore the mostly the background. For this reason we limit white portion remove from the patches so that it only contains the object regions in maximum possible way.

2.2.3 Creating tissue mask

Original Images have some background region and in the lumen area have same color as background have. So we create tissue mask for the object. This will be used to get the pixel level probability maps at the tissue regions only.

1. First read the image as grayscale images.
2. Apply Gaussian Filtering to remove noise.
3. Apply Otsu thresholding technique to assign the background as 0 and tissue as 1.
4. Add Padding to avoid weird border afterwards.
5. Use Dilation techniques to fill back holes since these are the lumen area that is in the object region.
6. After that use the erosion technique to restore the borders, eating up small objects.
7. Dilate again.
8. Crop to restore the original image with the tissue region filled with green colors.

2.3 Fine tuning the Deep CNN

After the Preprocessing step we go to fine tune Deep CNN architecture in the following steps:

1. Importing the necessary libraries(os, sys, pickle, glob, numpy, pandas)
2. From keras import mobile net architecture and corresponding preprocessing import keras.appications.imageutillls import the preprocess input. Here Mobile net is work in range[-1,1], so we appropriately preprocess to input the images.
3. From keras layers and utills import model, Dense ModelCheckpoint, Reduce learning rate On Plateu and balanced generator. This balanced usually augment the patched images as input.

4. Get the filenames stored in the info directory to get primary and secondary grade with images.

5. Initial Dimension is taken as 250x250 and dim for the architecture input is taken as 224x224.

6. For the training we have taken names with starting ZT199, ZT204, ZT111 and collect the gleason score from the corresponding csv files.

7. For validation set We use images naming ZT76 and collect the corresponding gleason score from the csv files.

8. Define the data generators with the training and validation patches

9. Use Mobile Net architecture with alpha value (width multiplier) 0.5. Change the model architecture at the top level by adding softmax activation with 4 classes as output.

10. Firstly train only the top layers which are randomly initialized and freeze all convolution layers with a Adam with learning rate 0.001 as optimizers and loss as categorical cross entropy loss for 500 iterations.

11. Then the all the layers train for 50000 iterations with SGD as as a optimizers with learning rate 0.001 and loss as categorical cross entropy.

12. Best model weights will be getting by saving the history and finding the weights for whom validation loss is minimum.

It is relatively small dataset(641 patients), so we find strong regularizations and balanced mini batches are to be crucial for successfully train the classifier.

2.3.1 Plotting Heatmaps and Class Activation mappings

After Fine tuning the network we get the best model weights using this we plot the heatmaps and class-activation mappings. Algorithms for this is as follows:

1. importing the necessary libraries(os, sys, glob, numpy, pandas, matplotlib, PIL, keras, from keras_utils preprocess_input_tf and center_crop).

2. Initial dim and dimension is taken as 250 and 224 respectively. Class labels as usual 4 classes 'benign'(Index 0), 'gleason3'(Index 1),'gleason4'(Index 2), 'gleason5'(Index 3) with previous color palettes.

3. Get the patch directory of TMA 80 images and also the masked by two pathologists and tissue masked images.

4. Load the csv file containing Gleason scores with filenames.

5. Load the trained patch level model with best weights.

6. For output pixel-level heatmaps, we first create a model for predicting on whole TMAs with rescaling factor 3.

7. Instead of Global pooling we use average pooling.

8. For class activation maps, we first image transformed to array, then crop center with size 1024 and then prepare for the network to visualize the class activation map.

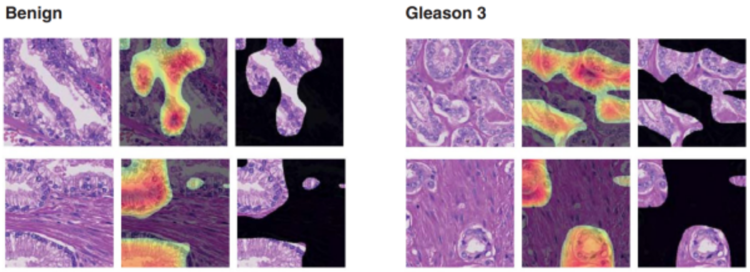


Figure 2.5

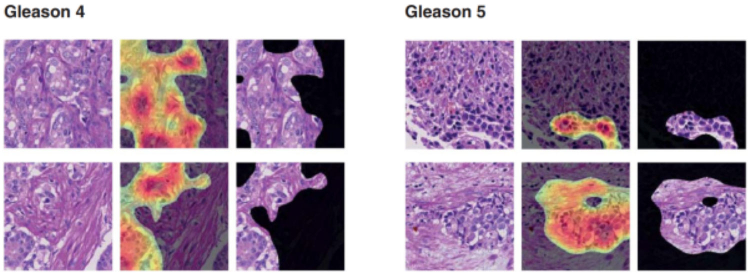


Figure 2.6

2.3.2 Plotting the test results

Finally we want to compare the model performance with pathologist level performance with confusion matrix.

2.4 Model Architecture

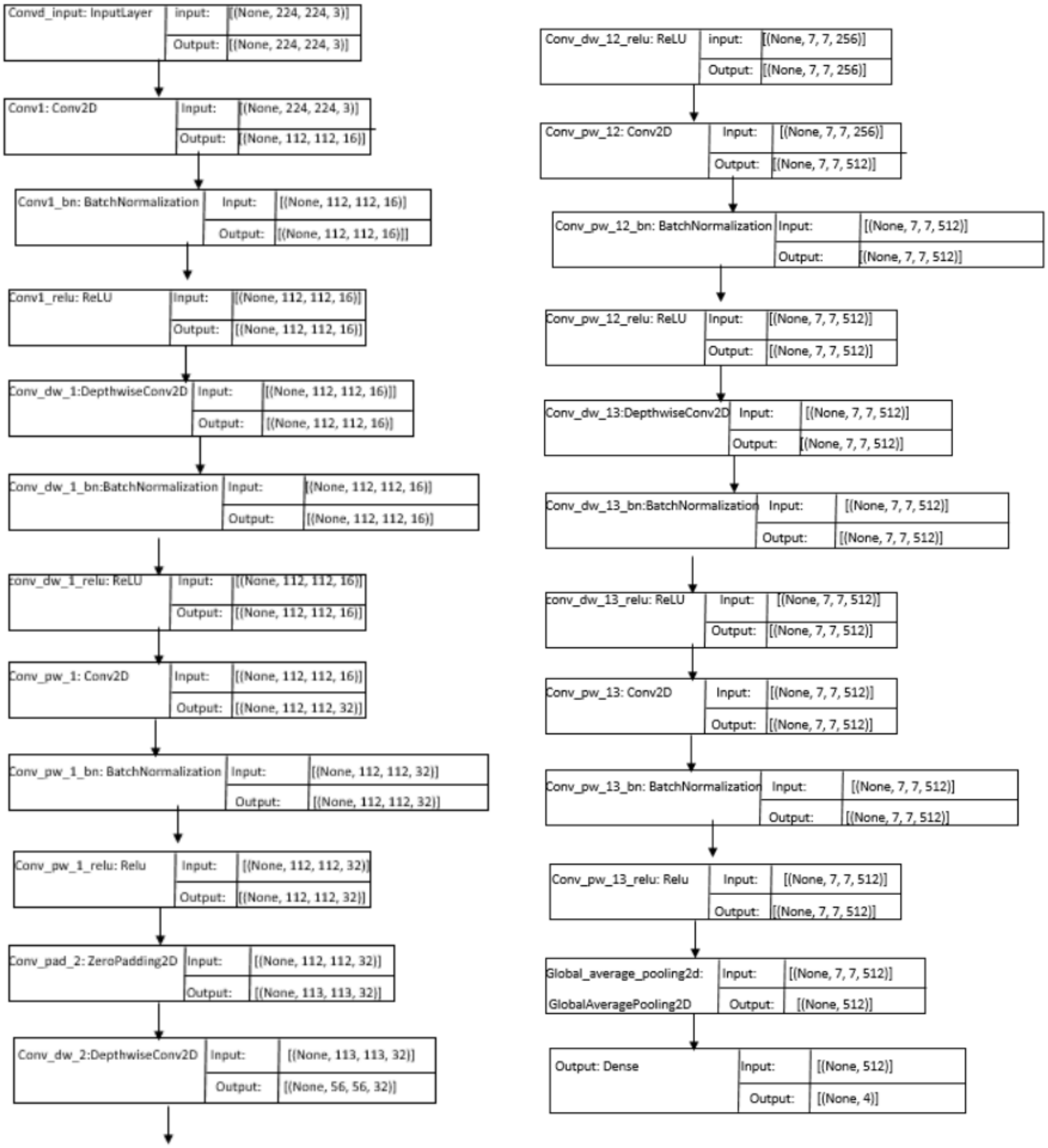


Figure 2.7: Trained Model(Mobile Net Architecture with $\alpha = 0.5$

Here is the Model architecture of the trained model. We use Mobile Net architecture with width multiplier $\alpha = 0.5$. Mobilenets have an extra property of depthwise separable convolutions in place of standard convolutions used in earlier architecture. Thus we have a lighter architecture meaning less number of parameters. In this paper we use 8.2 lakh trainable parameters.

CHAPTER 3

Results

3.1 Tissue Micro array resource with Gleason score annotation

This dataset have 5 tissue micro arrays(TMAs), each containing 200-300 spots. There are some artifacts present in the tissue these were been neglected from the study. TMA spots were annotated by a first pathologist classifying a Gleason pattern of 3, 4 or 5 to each region. These are shown in the below.

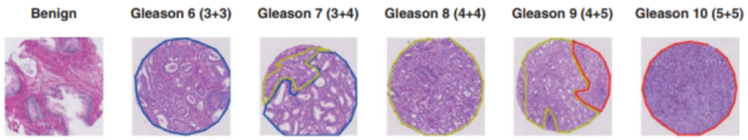


Figure 3.1: Glesaon Score annotation by pathologists

3.2 Automatic Gleason score annotation via Deep Learning

Small image patches were extracted from tissue regions and used to train a patch based classifier(Fig). After training the patch based classifier converted to pixel level annotator. Then it assigns Gleason scores to entire TMA spot images. Here we use Mobile Net architecture[15] with width multiplier $\alpha = 0.5$ which have shown great performance on imagenet competition. Mobilenets are designed for small models to be trained on mobile devices. Since our dataset is small so there is relatively low number of parameters would be required.

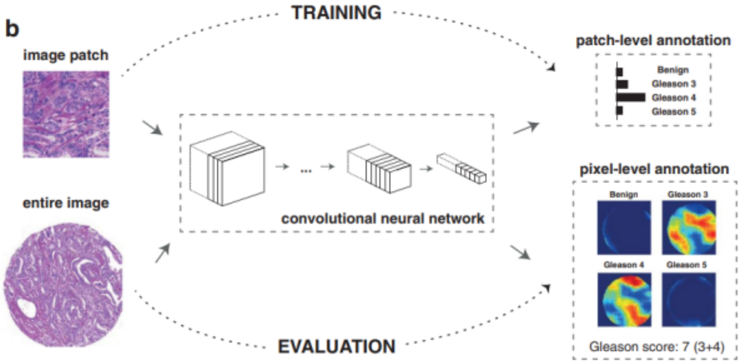


Figure 3.2: Training and evaluation phase

3.3 Classification Performance

We plot the confusion matrix for the validation dataset TMA ZT76. There is a measure called Cohen's quadratic kappa score[16] which shows the performance of the model due to interpathologist variation. Kappas score is 0 means the model is agreeing with the original by chance and 1 means model is perfectly agreeing. For ordered classes, weighted Cohen's Kappa is more important since it penalizes more strongly the inter annotator disagreement. The quadratic weighted kappa defined as follows:

$$kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \quad w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

where N = Total number of considered classes and the i, j indices refer to the ordered rating scores $1 \leq i, j \leq N$. $O_{i.j}$ denote the number of images that received rating score i by the first expert and rating score j from the second and $E_{i,j}$ denotes the expected number of images receiving rating i by the first expert and rating j by the second, assuming no correlation between rating scores. For validation set the kappa score is 0.67. Comparing to the interpathologist agreement kappa score is 0.71 which is close. Confusion matrix for the validation patches is given in the below figure.

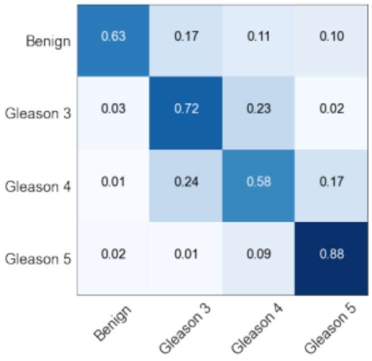
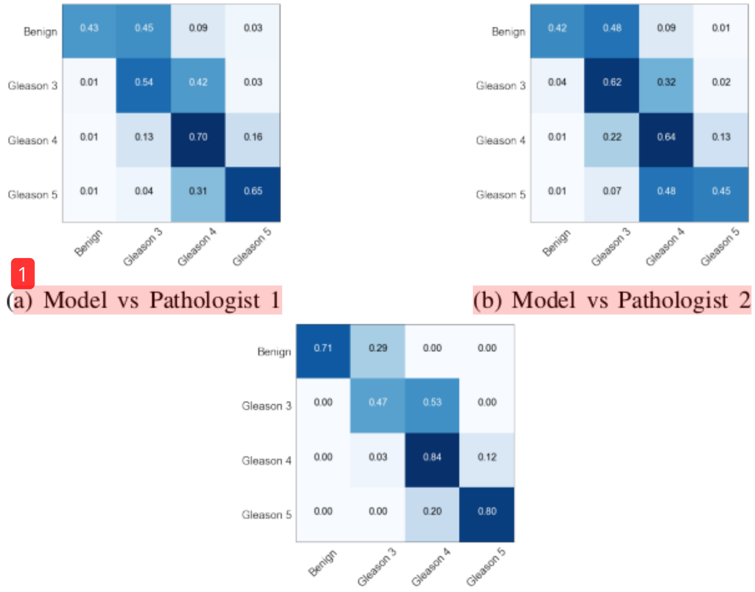


Figure 3.3: Confusion matrix for validation data

In the test patches we want to predict the model vs both pathologist1 and pathologist 2. Also there is a interpathologist variability confusion matrix.



(a) Model vs Pathologist 1

(b) Model vs Pathologist 2

(c) Pathologist 1 vs Pathologist 2

Figure 3.4: Confusion matrix for Gleason grade on patches

Finally we count the Gleason scores which is available in csv file and check for the validation data TMA ZT76. there are 6 different classes will be which are benign, gleason 6, gleason 7, gleason 8, gleason 9, gleason 10.

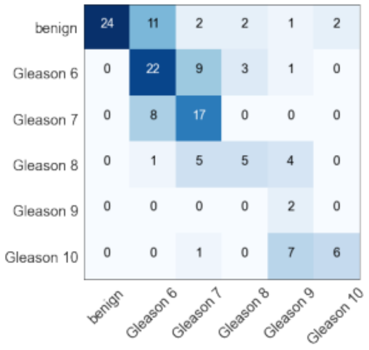
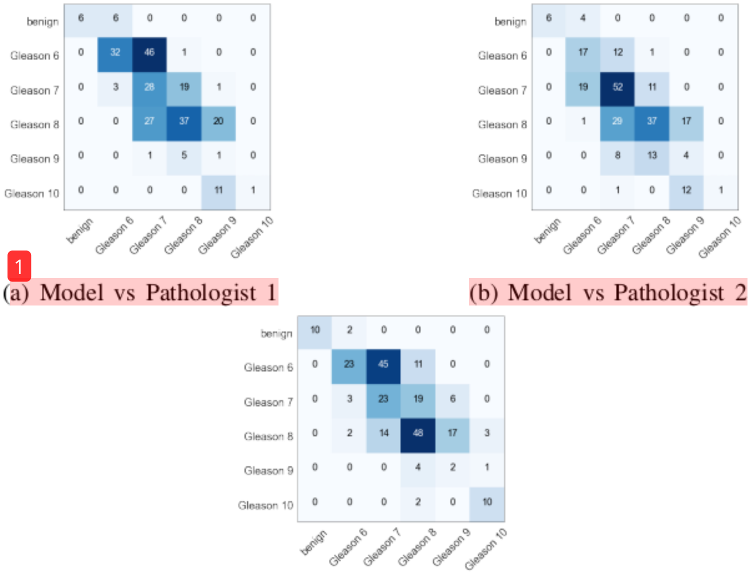


Figure 3.5: Confusion matrix for validation data

For the test cohort the cohen kappa value between model vs pathologist 1 is 0.75, model vs pathologist 2 is 0.71 and the inter-pathologist kappa score is 0.71. So we can see that our model more less predict as good kappa score near the value of inter-pathologist agreement.



(a) Model vs Pathologist 1

(b) Model vs Pathologist 2

(c) Pathologist 1 vs Pathologist 2

Figure 3.6: Confusion matrix for Gleason score

CHAPTER 4

Conclusion

In this work, we trained a DEEP CNN (Mobile net) as Gleason score annotator. Using this model we can predict Gleason score which is the most valuable predictor for the treatment purpose. By the Class activation Mappings we can visually predict the morphological patterns.

4.1 Limitations and Future work

We can see from the accuracy score which is below 60%. So the Gleason score prediction is not an easy way as we see it also varies between different pathologist. From the confusion matrix on Gleason score assignments we can see miss-classifications happens only for the nearby gleason score and model wrongly predict the class Gleason 10. If we take more pathologist from different hospitals then the result will be more accurate. Our studies are based on H&E stained Tissue micro array cores only. But it clinically correct to take needle core biopsies. Also this tissue images are taken only at a particular time. In future these tissue may be degraded and form cancer. So we have to take a timely evaluation of Tissue images to check how these tissue involves into degradation process.

APPENDIX A

CODE ATTACHMENTS

A.1 CREATING PATCHES

```
create_patches.py
```

A.2 CREATE TISSUE MASKS

```
create_tissue_masks.py
```

A.3 FINE TUNING THE CNN

```
gleason_score_finetune.py
```

A.4 PLOTTING HEATMAPS

```
plot_heatmaps_and_CAM.py
```

A.5 PLOTTING RESULTS

```
plot_test_cohort_results.ipynb
```

REFERENCES

[1] Long-Sheng Chen, Fei-Hao Hsu, Mu-Chen Chen, and Yuan-Chia Hsu. “Developing recommender systems with the consideration of product profitability for sellers”. In: *Information Sciences* 178.4 (2008), pp. 1032–1048.

[2] J. Epstein. “Prostate Biopsy Interpretation”. In: *Philadelphia, PA: Lippincott-Raven* 2nd ed. (1995).

[3] D.F. Gleason and G.T. Mellinger. “Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging”. In: *Journal of Urology* 111.1 (1974), pp. 58–64.

[4] Donald F. Gleason. “Histologic grading of prostate cancer: A perspective”. In: *Human Pathology* 23.3 (1992), pp. 273–279.

[5] M.B. Amin, D. Grignon, P.A. Humphrey, and J.R. Srigley. *Gleason Grading of Prostate Cancer: A Contemporary approach*. Urology. Lippincott Williams & Wilkins, 2003. ISBN: 978-0781742795.

[6] Jonathan I. Epstein. “Prostate cancer grading: a decade after the 2005 modified system”. In: *Mod. Pathol.* (2018), pp. 47–63.

[7] Thomas J. Fuchs and Joachim M. Buhmann. “Computational pathology: challenges and promises for tissue analysis”. In: *Comput Med Imaging Graph* 35.7-8 (2011), pp. 515–530.

[8] J. Diamond, N. Anderson, P. Bartels, R. Montironi, and P. Hamilton. “The Use of Morphological Characteristics and Texture Analysis in the Identification of Tissue Composition in Prostatic Neoplasia”. In: *Human Pathology* 35 (2004), pp. 1121–1131.

[9] R. Stotzka, P.H. Bartels, and D. Thompson. “A hybrid neural and statistical classifier system for histopathologic grading of prostate lesions”. In: *Anal. Quant. Cytol. Histology* 17 (1995), pp. 204–218.

[10] A.W. Wetzel, R. Crowley, S.J. Kim, R. Dawson, L. Zheng, Y.M. Joo, Y. Yagi, J. Gilbertson, C. Gadd, D.W. Deerfield, and M.J. Becich. “Evaluation of prostate tumor grades by content-based image retrieval”. In: *Proc. SPIE AIPR Workshop Advances Computer-Assisted Recognition, Washington, DC.* Vol. 3584. 1999, pp. 244–252.

[11] Ali Tabesh, Mikhail Teverovskiy, Ho-Yuen Pang, Vinay P. Kumar, David Verbel, Angeliki Kotsianti, and Olivier Saidi. “Multifeature Prostate Cancer Diagnosis and Gleason Grading of Histological Images”. In: *IEEE Transactions on Medical Imaging* 26 (2007), pp. 1366–1378.

[12] “LeCun, Y. and Bengio, Y. and Hinton, G.” In: *Nature* 521 (2015), pp. 436–444.

[13] N. Zhou, F. Fedorov A.and Fennessy, R. Kikinis, and Y Gao. “Towards grading gleason score using generically trained deep convolutional neural networks”. In: *IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1705.02678 (2017), pp. 1163–1167.

[14] H. Kallen, J. Molin, A. Heyden, C. Lundstrom, and K. Astrom. “Large scale digital prostate pathology image analysis combining feature extraction and deep neural network”. In: *arXiv preprint* 1705.02678 (2017).

[15] A.G. et al. Howard. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *arXiv preprint* arXiv:1704.04861 (2017).

[16] J. A Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Education and Psychological Measurement* 20.1 (1960), pp. 37–46.

Prostate cancer diagnosis and gleason grading using histological images

ORIGINALITY REPORT

17%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.research-collection.ethz.ch Internet	147 words —	4%
2	www.ijsrp.org Internet	130 words —	3%
3	www-spacl.ece.arizona.edu Internet	97 words —	3%
4	v1.overleaf.com Internet	90 words —	2%
5	library.isical.ac.in:8080 Internet	35 words —	1%
6	www.biorxiv.org Internet	35 words —	1%
7	www.ncbi.nlm.nih.gov Internet	26 words —	1%
8	mdrl.mne.psu.edu Internet	21 words —	1%
9	mafiadoc.com Internet	19 words —	1%

10	www.nature.com Internet	16 words — < 1%
11	Peng, Yahui. "Computer-aided histological analysis for prostate cancer diagnosis", Proquest, 20111003 ProQuest	15 words — < 1%
12	www.freepatentsonline.com Internet	14 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 14 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 14 WORDS