CNN for Prostate Cancer Diagnosis and Gleason Grading of H&E Images.

Stathis Megas[1], David Zheng[2]

[1]UCLA, Physics and Astronomy
[2]UCLA, Medical Informatics

Abstract. Prostate cancer is the most common form of cancer in males in North America and has the third highest incidence among all cancers overall. Digital pathology techniques utilizing intelligent deep learning models for classifying disease are relatively new and have demonstrated potential for vastly improving both the speed and accuracy of diagnoses. In this report, we aim to construct deep learning models based on two different architectures – Residual Networks (ResNets) and DenseNets – for the purpose of classifying microscopy images of prostate cancer as either benign or malignant. Both architectures were then also used to build and train additional models with the goal to classify slides based on cancer grade and Gleason score. The best results obtained were a test accuracy of 0.9325 for the binary classification task and an accuracy of 0.617 for the cancer grading task, with both results having been based on DenseNets. This study successfully replicated attempts at binary classification which already appear in literature but was less successful at the grading task due to technical and time constraints.

Keywords. Prostate Cancer, Classification, Machine Learning, Deep Learning, Residual Networks, DenseNets, Gleason Score.

1 Introduction

Prostate cancer consistently ranks as one of the types of cancer with the highest incidence and mortality [1]. In 2020, prostate cancer was the type of cancer with the third highest incidence but fifth highest mortality in the country. The relatively lower mortality rate compared to incidence is due to effective treatment options being available when the disease is diagnosed early. In fact, the five-year survival rate for patients diagnosed with prostate cancer is as high as 98%[2]. This emphasizes the importance of early monitoring and detection of prostate cancer.

Currently, the four most commonly used methods for diagnosing prostate cancer are the prostate-specific antigen (PSA) test, magnetic resonance imaging (MRI), trans-rectal ultrasound (TRUS), and needle biopsy, with needle biopsies considered the gold standard to establish a diagnosis of prostatic adenocarcinoma [3]. Diagnosis of prostate cancer is predominantly based on microscopic criteria and only to a secondary degree on the clinical history. Some of the most reliable such microscopic criteria are infiltrative growth pattern, prominent nucleoli and lack of basal cells. However, recognizing these features in tiles with H&E staining is often challenging for pathologists due to the highly complicated nature of these qualitative criteria and the subtle nature of features characteristic to the different grades of prostate cancer.

The detection of the presence and grade of cancer along with other clinical factors are fundamental for medical decision making and determining the correct course of treatment. Low risk disease may be successfully treated with surgery, and in many cases the course of action is to simply undergo active surveillance if the risk is deemed too high due to other factors. Patients with prostate cancer categorized at higher risk may need to consider hormone or radiotherapy in conjunction with surgery in order to resolve their disease.

Detection of the qualitative criteria that distinguish cancer vs non-cancer and the different grades of cancer in imaging creates an opportunity for Deep Neural Networks (DNN) which have been able to achieve promising results in the detection of such features [4]. Their success relies on the use of convolutional layers that are able to extract features as part of the training process, thereby quantifying the qualitative criteria. Using computer assisted tools to help automate the diagnosis of prostate cancer offers several advantages. Besides reducing the burden on human resources, machine learning driven detection systems for prostate cancer may also lead to higher accuracy, earlier detection, and potentially higher survival rates for patients.

2 Methods

In this section, we begin by describing the data collection method. After that, we formally define our problem in the form of standard classification problems. Then, we describe the different frameworks that we used to solve our problem, and finally, we provide evaluation metrics on which our model was assessed and compared with previous efforts.

2.1 Data Description

Our dataset consists of four sets of images of needle biopsies of the prostate with H&E staining, and all of them were collected from the archives of the Pathology Department at Cedars-Sinai Medical Center. These images are at low-power magnification and were provided to us after a downsampling from 1200x1200 pixel resolution to 512x512. The first set (Set A) includes 1000 images from 1000 different patients, evenly split so that 500 have been labelled (by our gold standard which we explain in the next paragraph) as cancerous and 500 labelled benign. Set B has 300 different patients and 10 different images from each patient. For every patient, all of their 10 images have the same label, which is 0, 1, or 2 to denote respectively benign, low grade and high grade prostate cancer in the Gleason pattern classification scheme. Set B is also balanced since there are 1000 images for each of the 0,1,2 classes. Set C is the normalized version of Set A; and Set D the normalized version of Set B. Sets C and D were generated after it was noticed that a ‘vanilla’ mean pixel normalization, was negatively impacting images with a lot of empty space, which shows up (bright) white. So Sets C and D have normalized only the meaningful part of the images, which helped get rid of uneven staining between photos.

The gold standard used for the labelling just described is the algorithm described in [6] which has been shown to diagnose PCa from needle biopsy images with high performance in several metrics. The algorithm in [6] takes 100 images per patient and can diagnose the presence or absence of cancer, and if cancer is present, whether it has low grade or high grade differentiation. For Set A (and C) the images provided to us were the images that got the highest attention from the algorithm in making the determination of the diagnosis. Similarly, for Set B (and D) we were given from each patient the 10 images with the highest attention among 100 images, so presumably, all 10 of them are exhibiting symptoms clearly indicative of their assigned label 0 through 2. The algorithm in [6] was developed by supervised learning on images hand-annotated by an expert research pathologist which were then cross-evaluated by other pathologists, and corrections made by consensus.

Our dataset was split into test, validation and training according to the fractions 20%, 20%, 60%, and, using the high-quality ANTIALIAS tool, we tried several different downsamplings of the images ranging from 64x64 to the original size 512x512.

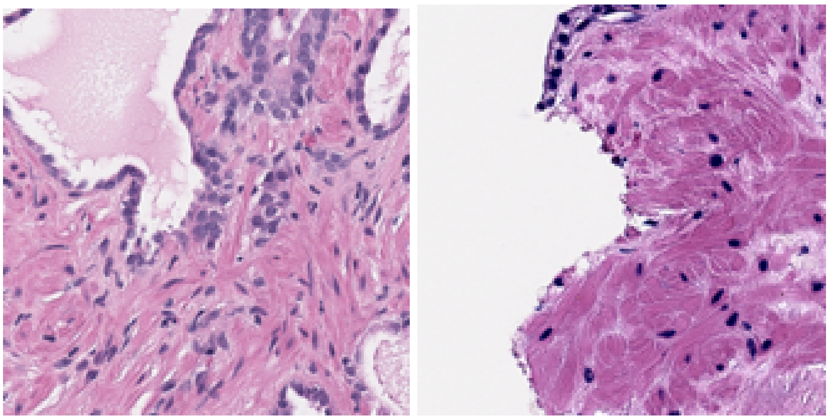

Fig. 1. On the left, an example image from Set A (unnormalized), and on the right, an example from Set C (normalized).

2.2 Model description

The cancer diagnosis problem and the Gleason grading problem were treated as a binary classification and three-class classification problem, respectively. Although the cancer diagnosis problem is a sub-case of the three-class classification problem that includes information about the grading, we chose to train different neural networks to achieve high performance on these two separate tasks. In the future these different networks could be used together to make ensemble predictions.

Both of these classification problems are machine vision problems, and there is nowadays a vast range of networks that have shown great performance at image recognition. In this study, we examined the architectures of ResNets and DenseNets which used different novel ideas to revolutionize machine vision. Both architectures are Convolutional Neural Networks, and as such they draw their inspiration from the pyramidal structure of cells based in the cerebral cortex. The main observation that led to the invention of ResNets is that even when using batch-normalization, which to some extent addresses the vanishing gradients problem, shallower traditional CNN achieved smaller train loss than their deeper counterparts. In other words, traditional CNNs have trouble learning the identity function, and hence in [10] they introduced skip (identity) connections between the layers to assist the learning of the identity function, which indeed significantly increased the manageable depth of CNNs. The central idea behind DenseNets is to keep the idea of skip connections, but instead of just adding the value of skip connection to the output of their destination layer, DenseNets concatenate the value of the skip connection with the output of the destination layer [9].

A final important decision was whether to implement our algorithm and analysis locally or in a computation cloud. We chose to try both approaches so as to gauge whether the two tasks pursued in this work could adequately be implemented in a simple computing cloud like Google Colab.

2.3 Statistical evaluation

Our datasets were chosen to be balanced for two important reasons: to reduce the probability of local minima in the loss function that read to a uniform benign prediction from the algorithm, and expand the metrics available to test performance. Performance metrics can be broadly divided into two classes: the ones which depend only on the performance of the test/algorithm, and the ones that depend on both the performance of the test and the prevalence of the disease in the population. An example of the latter is the accuracy, defined as

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}.$$

An example of the former category is the confusion matrix, which for a two-class classification reads

	predicted P	predicted N
actual P	TP	FN
actual N	FP	TN

However, when using balanced datasets as we do in this work, both kinds of metrics offer valuable insight. Hence, we chose to use the accuracy as the main performance metric for the task of cancer detection, and the confusion matrix for the task of Gleason grade classification.

3 Results

In this section we present the findings of our experiments and analysis. We do so separately for each of the local and the remote implementations.

3.1 Remote implementation on Google Colab

3.1.1 Binary classification task

For the binary classification of the images in Set A into benign or cancerous, we used the ResNet56v2 architecture and performed experiments to identify the downsampled resolution that would lead to the best performance (Fig. 2). We observed the highest accuracy for the 128x128 resolution however it was only roughly 2σ higher than the accuracy for 256x256. Since theoretically an ideal network would perform better for images of higher resolution, we kept using the 256x256 resolution.

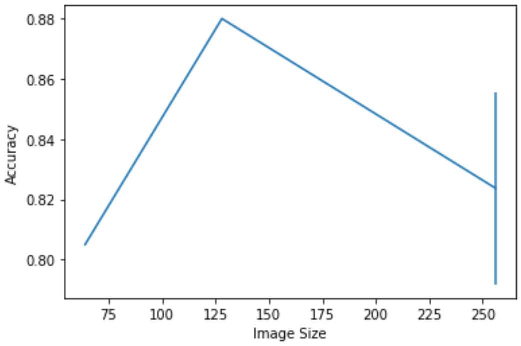

Fig. 2. Accuracy on the test set for the models with the best validation accuracy achieved on images with resolution 64x64, 128x128, 256x256. The error bar plotted for the accuracy of the resolution 256x256 has been derived as the MSE of the test accuracies when we vary the train and validation subsets. We thought this is a measure of the effect of our choice of the training set.

A manual err analysis showed that the algorithm persistently incorrectly classifies specific slides with high softmax probability. Indeed a t-SNE analysis for the high resolution clearly indicates some outliers (Fig. 3).

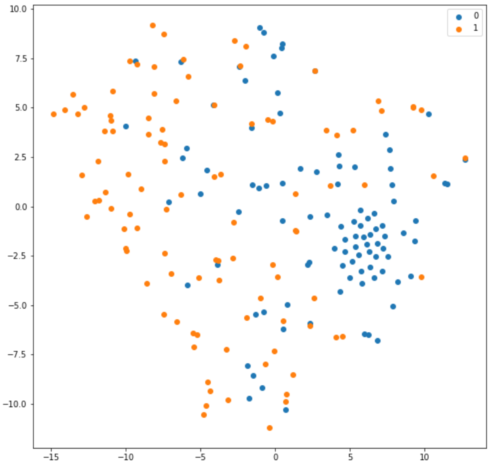

Fig. 3. On the left, t-SNE dimensional reduction (with PCA preprocessing) were points are plotted according to their real class: 0 for benign, 1 for cancerous. A few of the persistently misclassified images have a lot of empty space.

3.1.2 Three-class classification task

For the Gleason grading classification into benign, low grade and high grade classes, we used Set D, and we found the best results for the DenseNet201 architecture. Since the images D has many photos from each patient we performed an analysis both for tile and patient classification (Fig. 4,5). For the patient level classification we aggregated the results from the tiles by predicting the classes with the biggest softmax probability after averaging over the 10 images of each patient.

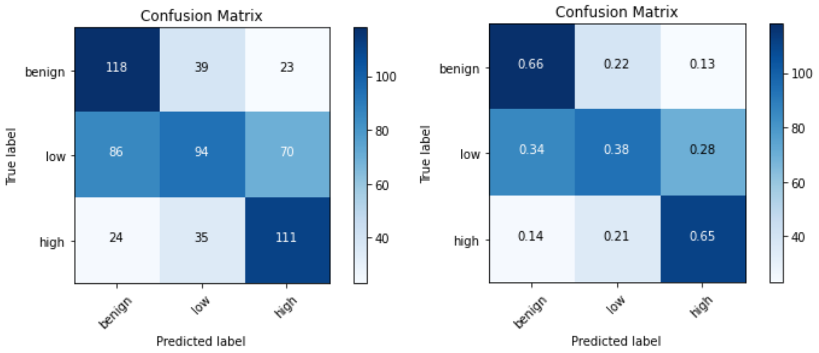

Fig. 4. Confusion matrix (normalized on the left, and un-normalized on the right) for tile classification into benign, low grade and high grade.

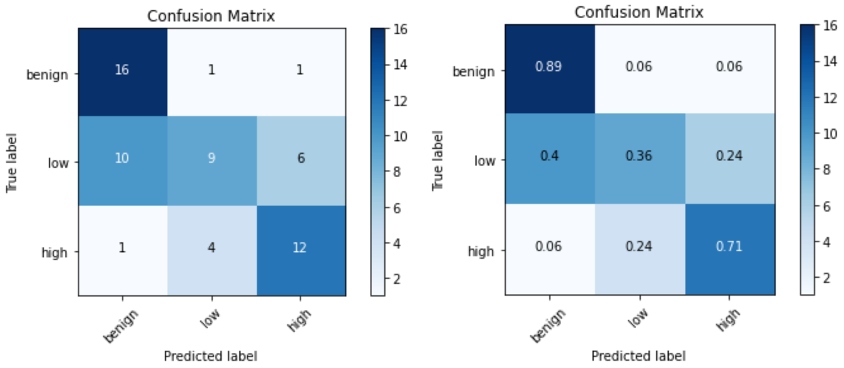

Fig. 5. Confusion matrix (normalized on the left, and un-normalized on the right) for patient classification into benign, low grade and high grade.

We observe that the biggest values of these four matrices are concentrated along the main diagonals and that the accuracy improves at the patient level, as expected. At the tile level the binary accuracy is 0.538, but at the patient level it improves to 0.617. Our model is also very prone to misclassifying the low grade into the other two classes, indeed this shows up also in the t-SNE dimensional reduction (Fig. 6)

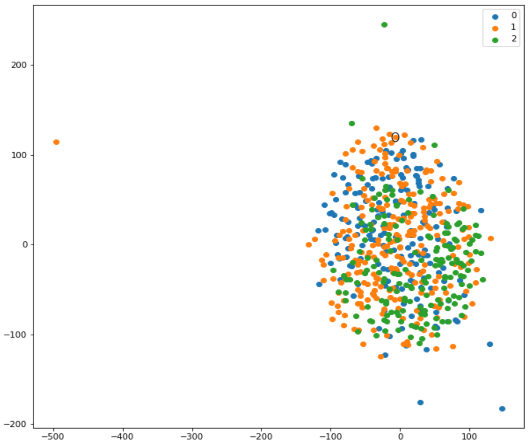

Fig. 6. t-SNE analysis of the final concatenation layer of the DenseNet201. There is good separation between benign (0) and high grade (2), but low grade (1) is prone to being misclassified.

3.2 Locally Constructed Models

The locally constructed models were trained and evaluated on a machine with 16gb RAM and an Nvidia RTX 3090 with 24gb VRAM. The locally constructed Resnet used the Adam optimizer with a learning rate starting at 0.001 and decreasing by a factor of 10 every 10 epochs for a total of 50 epochs. The Resnet code was heavily based upon an online tutorial which used ResNet50 as its reference. It was able to achieve a test accuracy of 0.835 with an AUC of 0.847 on the dataset A (1000 images used for simple classification of benign vs malignant).

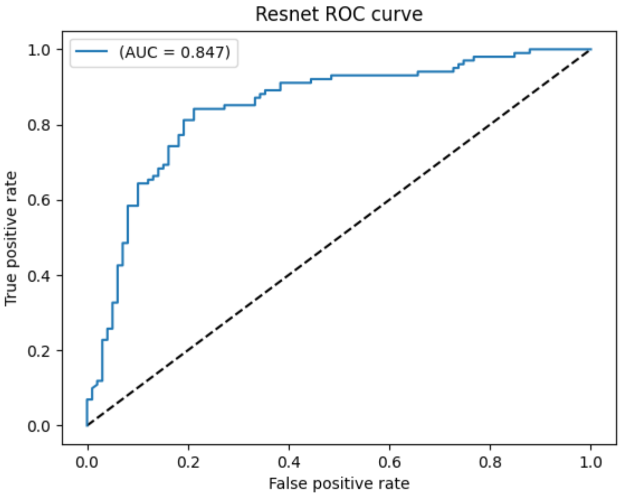

Fig. 7 ROC curve for the Resnet- trained and evaluated on Set A.

The constructed Densenet, based on the template DenseNet121 Keras architecture implementation, as its reference, was able to achieve a test accuracy of 0.95 with an AUC of 0.988 on Set A. We used the same learning rate curve as the previously mentioned Resnet with the optimizer being changed to SGD.

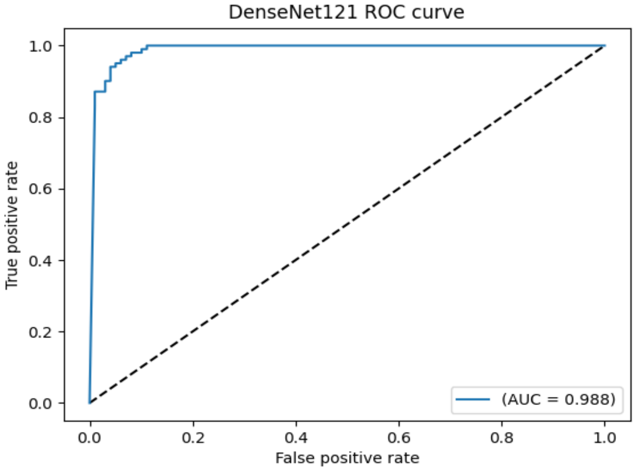

Fig 8. ROC curve for the densenet - trained and evaluated on Set A.

The same models were trained and evaluated on a subset of Set C, which had the 1000 images indicating a benign sample removed. The motivation for this was to see whether or not the models would be able to distinguish low-grade prostate cancer from high-grade prostate cancer. Thus, the models in this case were only trained on 2000 images, 1000 from low-grade and 1000 from high-grade.

The resnet was able to obtain a test accuracy of 0.8275 with an AUC of 0.898 while the densenet was able to obtain a test accuracy of 0.9325 with an AUC of 0.982.

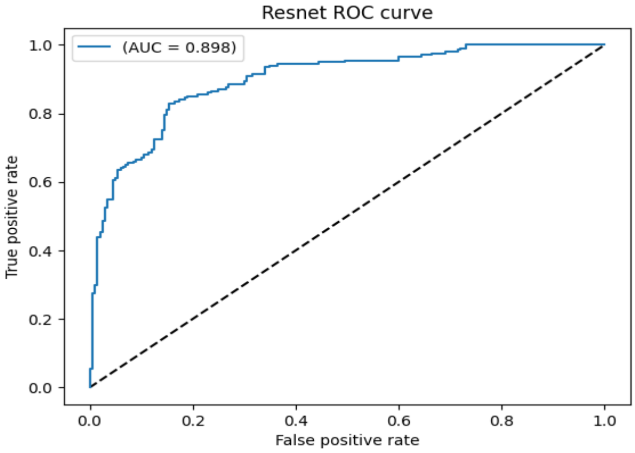

Fig 9. ROC curve for the resnet - trained and evaluated on a subset of Set C

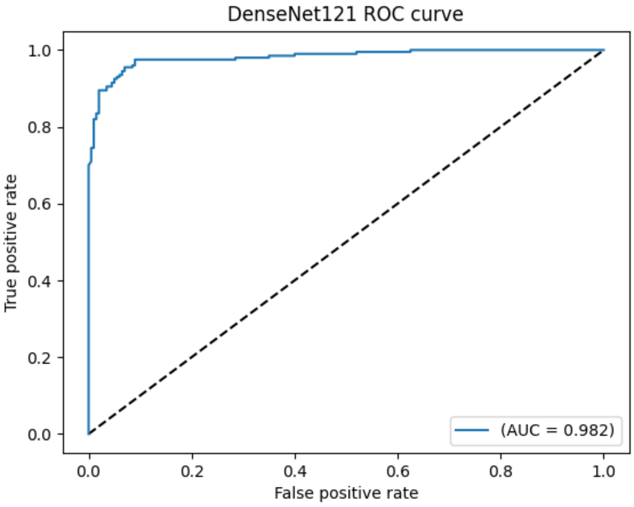

Fig 10. ROC curve for the densenet - trained and evaluated on a subset of Set C

There were also attempts made to train both classifiers on the entirety of Set C, however, these attempts met little success. Accuracies achieved did not amount to higher than just random guessing, and fine-tuning the models was complicated by technical difficulties and memory limitations possibly caused by inefficiencies in the code along with time limitations.

A cursory glance at the ROC curves and accuracies produced by both models across both trails appears to indicate higher performance for the densenets. This suspicion was confirmed through the construction of contingency matrices based off the test results and then performing McNemar's test.

	Densenet Correct	Densenet Incorrect	Total
Resnet Correct	161	6	167
Resnet Incorrect	29	4	33
Total	190	10	200

Fig 11. Contingency Table for models on Set A

	Densenet Correct	Densenet Incorrect	Total
Resnet Correct	319	12	331
Resnet Incorrect	54	15	69
Total	373	27	400

Fig 12. Contingency Table for models on subset of Set C

The McNemar's test p-values were $P = 1.2*10^{-5}$ for the models trained on Set A and $P = 1.7*10^{-7}$ for the models trained on the described subset of Set C. The low p-values ($P < 0.05$) indicates that the models are statistically different. Due to the higher accuracy demonstrated by the densenet in both cases, we claim that densenet is the superior deep learning architecture for classifying benign vs malignant prostate slides as well as low-grade vs. high grade.

4 Discussion

In this section, we start with comparisons of the local and remote implementations with baseline models from the literature, and then proceed to explain limitations of our method.

4.1 Model Comparisons

Comparison between the local and remote implementations shows that the local implementation performed significantly better in the binary classification task, with a test accuracy of 0.9325 and an AUC of 0.982. This is partly due to bigger memory and usage constraints inherent in the remote implementation. However, the remote method seemed to give better results in Gleason grade classification where, at the patient-level, a test accuracy of 0.617 was achieved. Moreover, if we used the grade-classification model for the benign-malignant classification, the accuracy would be 0.783, at the patient level. It is also interesting to note that both these independent implementations were ultimately led to use DenseNet architectures over ResNet, for both tasks. However, this doesn't necessarily mean that DenseNets are more suitable for the structure present in H&E stain images, since DenseNets outperform ResNets in CIFAR competitions too.

The results obtained in the binary classification also favorably compare to models that have already been explored in the literature. In the recent review [4], most of the slice level AUC reported are in the range of 0.8-0.9. However, in [6] much greater performance was reported for the segmentation task, which is usually a harder task. For the Gleason-grade classification, the accuracy we achieved is probably low compared to the literature. Most papers published, e.g. [7,8], looked at the classification into low vs high, so that their results are not directly comparable to ours. For instance, at [7] which performed a Gleason grade classification into 5 cancerous classes we see in their Fig. 4 that the induced low vs high grade classification they achieve is 0.781.

4.2 Limitations and Future Directions

The comparison with benchmark models from the literature also reveals our work's limitations and possible path for improvement. In the binary classification, there seems to be small possibility of improvement, but for the Gleason grading there is vast possibility of improvement, both in terms of trying different architectures for a three-class model, as well as in combining our two models. For instance one idea that to explore in the future is implementing a pipeline that first classifies images in terms of benign vs malignant, and then classifies the malignant into low vs high grade. This would make use of our binary model's excellent performance to eliminate the most important error in the confusion matrix in Fig. 4,5, namely low grade cancer being classified as benign. Indeed this seems to be the approach followed in the literature since most papers, e.g. [7,8], seem to only attempt classifications of Gleason grading of cancer separately from the benign-malignant classification.

5 Conclusion

While it is of the authors' opinions that the models produced in this study are not ready for usage in clinical settings, the promising results indicate that deep learning is indeed a viable path to pursue for the diagnosis and grading of prostate cancer through microscopy images. While the performances of the models used for grading were relatively poor, the high accuracies achieved locally regarding binary classification between benign vs. malignant and low-grade vs. high-grade indicate that deep learning is indeed suited for the task of grading – better results will be possible given additional time and fine-tuning of the models. Application of computer-aided diagnosis and grading of prostate cancer would increase the speed and accuracies of diagnosis – leading to faster and more suitable treatments for the disease. If it is applied successfully, the integration of deep learning into diagnostic procedures will improve outcomes for patients and greatly assist physicians in the process of medical decision making.

Acknowledgements

Thanks goes to Dr. Jiayun Li for providing the slides used in this project. Additional thanks goes out to fellow students Joy Fu and Rina Ding for input and assistance with model construction. Finally, special thanks to Dr. Corey Arnold and Dr. William Speier for organizing weekly meetings as well as their helpful insights and general guidance throughout the quarter.

References

[1] Common cancer types : https://www.cancer.gov/types/common-cancers

[2]https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21590

[3] Prostate cancer: diagnostic criteria and role of immunohistochemistry, Cristina Magi-Galluzzi, https://www.nature.com/articles/modpathol2017139.

[4] Prostate cancer Detection using Deep convolutional neural networks, Sunghwan Yoo, Isha Gujrathi, Masoom A. Haider & farzad Khalvati, https://www.nature.com/articles/s41598-019-55972-4

[5]https://github.com/jeffheaton/t81_558_deep_learning/blob/master/t81_558_class_06_3_resnet.ipynb

[6] Li W, Li J, Sarma KV, Ho KC, Shen S, Knudsen BS, Gertych A, Arnold CW. Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images. IEEE Trans Med Imaging. 2019 Apr;38(4):945-954. doi: 10.1109/TMI.2018.2875868. Epub 2018 Oct 12. PMID: 30334752; PMCID: PMC6497079.

[7] Bejoy Abraham, Madhu S. Nair, Automated grading of prostate cancer using convolutional neural network and ordinal class classifier, Informatics in Medicine Unlocked, Volume 17, 2019, 100256, ISSN 2352-9148,https://doi.org/10.1016/j.imu.2019.100256.

[8] Karimi D, Nir G, Fazli L, Black PC, Goldenberg L, Salcudean SE. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images-Role of Multiscale Decision Aggregation and Data Augmentation. IEEE J Biomed Health Inform. 2020 May;24(5):1413-1426. doi: 10.1109/JBHI.2019.2944643. Epub 2019 Sep 30. PMID: 31567104.

[9] Densely Connected Convolutional Networks, Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, arXiv:1608.06993

[10] Deep Residual Learning for Image Recognition, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, arXiv:1512.03385