Automated Classification Map Generation of Prostate Cancer using Deep Learning

Wenhan Tan
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA, USA
wt99@drexel.edu

David E. Breen
Department of Computer Science
Drexel University
Philadelphia, PA, USA
david@cs.drexel.edu

Fernando U. Garcia
Department of Pathology
Tower Health
Reading, PA, USA
fernando.garcia@towerhealth.org

Mark D. Zarella
Department of Pathology
Johns Hopkins University
Baltimore, MD, USA
mark.zarella@jhu.edu

***Abstract*—Whole slide images are examined by pathologists and scored according to the Gleason grading system. It is a time-consuming task and may involve assessing variability between different pathologists. In this work, a deep learning system is presented that generates classification maps for whole slide images. This system produces patch-level results first and then predicts a classification map for each prostate cancer slide. The classification maps contain regional cancer severity for each biopsy and are compared with provided mask images. Both provided mask images and predicted mask images are then reviewed by an experienced pathologist to evaluate classification performance. Most state-of-the-art deep learning methods cannot explain how they output classification results. With this work's classification maps, pathologists can see the regional classification results that explain the algorithm's classification.**

Keywords—prostate cancer, Gleason grading system, whole slide image, deep learning, convolutional neural network, classification maps

I. Introduction

Prostate cancer is the most common cancer for males worldwide, leading to more than 350,000 deaths every year. Around 10% of men will be diagnosed with prostate cancer in their lifetimes and about 50% of them need to be actively treated. It is important to start treatment at an early stage to reduce the mortality rate of prostate cancer. The initial diagnostic responsibility often falls on a pathologist, who must determine the grade and severity of the cancer. Usually, pathologists examine tissue samples as whole slide images, which are microscopic level images of hematoxylin and eosin (H&E) stained tissue from a prostate tumor. Diagnosis of these whole slide images follows the Gleason grading system, which evaluates the structure, stage and aggressiveness of a prostate cancer sample. However, grading variability exists between different pathologists and may result in unnecessary or insufficient treatments. There are many factors that lead to grading variability among pathologists. The same tissue sample may have different appearances, depending on how the H&E staining and imaging are performed, as well as the type of scanner and processing used to acquire the image.

The ability to annotate regions of a histopathology slide in an automated fashion has several advantages. First it enables a pathologist or technician to examine the slide and find the most informative regions for Gleason grading, providing a tool to support human-in-the-loop applications of AI. This has the potential to reduce the likelihood of misdiagnosis from regions of the slide that may not have been viewed. Second, it can enable efficient three-dimensional segmentation strategies, in which multiple serial sections can be quickly and automatically annotated, producing three-dimensional volumes for volumetric analysis. Third, it can support other studies that rely on regional annotation which may be difficult to acquire at scale; for example, labeled regions of the histopathology image can be transferred to other imaging modalities in the same tissue, providing a rapid and objectively labeled ground truth for interpretation of new data sets.

Tissue samples must first be cut from patients, then scanned by medical scanners, and eventually displayed on a computer. These steps, along with examination by a pathologist, create a time-consuming process. Recently, deep learning has been seen as a powerful AI tool for healthcare and more specifically in the digital pathology field. These methods have been shown to produce good accuracy in slide-level classifications. Even experienced pathologists can only assess the severity of a prostate cancer sample with an accuracy of 80%, while some state-of-the-art deep learning models can achieve accuracy above 90%. Clinically, pathologists have seen accuracy increase when diagnosing cancer with help from AI tools. However, model development has not focused on patch-level labeling and classification map creation. The development of this kind of deep learning model is beneficial for labeling multi-center datasets and clinical practice usage.

The Gleason grading system outputs a final Gleason Score which indicates the two most prominent cancer patterns present in a tissue sample [1]. There are three Gleason scores labelled pattern 3, pattern 4, and pattern 5. An example of a Gleason score would be 3 + 4. The first number (3) is the most prominent Gleason pattern in the specimen and the second number (4) indicates the second most prominent pattern. There are also five Gleason groups named Grade Group 1 to 5. Table I explains the relationship between Gleason Scores and Gleason Groups. The higher the Gleason Group number is, the more malignant the cancer is, and the more aggressive the needed treatment is.

We present a deep learning system that generates classification maps based on the Gleason grading system for whole slide images [15]. This system produces patch-level results first and then predicts a classification for each prostate cancer slide. The classification maps contain regional cancer severity for each biopsy and are compared with provided mask images. Both provided mask images and predicted mask images are then reviewed by an experienced pathologist to evaluate classification performance. With this work's classification maps, pathologists are able to see the regional classification results that explain the algorithm's overall classification.

TABLE I. Relationships between Gleason groups and gleason score. Note that gleason score 3, 4, and 5 correspond to gleason pattern 3 ,4, and 5.

Gleason Group	**Gleason Score**
Group 1	3 + 3
Group 2	3 + 4
Group 3	4 + 3
Group 4	3 + 5, 4 + 4, 5 + 3
Group 5	4 + 5, 5 + 4, 5 + 5

II. Previous WOrk

A. Color Normalization & Data Augmentation

Whole slide images are obtained from microscope scanners that require precise control to clearly capture the cellular structure of tissues for a pathologist's examination. Frequently the tissues are stained before scanning. Hematoxylin and eosin (H&E) stain is the most widely used stain in medical diagnosis. The hematoxylin stains cell nuclei blue, and eosin stains the extracellular matrix and cytoplasm pink, with other structures taking on different shades, hues, and combinations of these colors. Other factors may affect the acquired images during scanning, for example, scanner brightness, distance between tissues and scanners, angle differences, etc. Differences in staining chemicals and procedures may also affect the appearance of the stained tissues. Example H&E scans are shown in Fig. 1. Notice that color variation exists even in the patches scanned by the same scanner from the same data center.

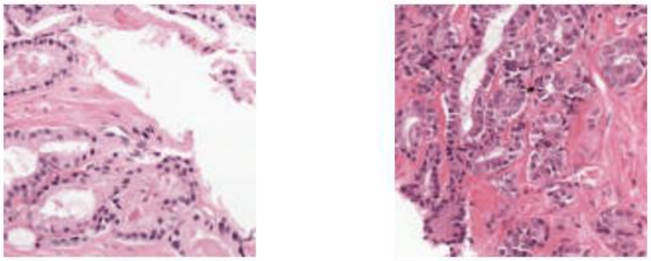


Fig. 1. Two example patches from the training set. They are both pattern 3 from the same scanner and data center. Color variations still exist between these two patches.

In the case of "learning" colors instead of tissue structures by deep learning models, the technique of color normalization is introduced to "normalize" the color space in whole slide images to guarantee that they have the same color. Many color normalization techniques have been developed over the years. Algorithms including histogram normalization [4], stain separation [5, 6], non-linear mapping based on spatial properties [7], and even generative adversarial network have been used for this data pre-processing step [8].

In addition, data augmentation has also been proposed to increase the data samples by generating similar images from the original training sets. These images are generated by randomly rotating the original images by 90 degrees, flipping them horizontally or vertically; to some more advanced techniques including Gaussian blurring and adding Gaussian noise. In fact, color normalization can be seen as a subset of data augmentation, since data augmentation can also randomly adjust brightness, hue, and saturation. This forces the model to focus on things like tissue structure, instead of colors. In Tellez et al. [9], their algorithms demonstrate the effectiveness of color normalization in image classification performance. Data augmentation or color normalization techniques should be applied since they usually increase performance.

B. Deep Learning in Prostate Cancer

In Bulten et al. [2], an automated deep learning system has been proposed to grade prostate cancer following the Gleason grading system. Data is collected from the Radboud University Medical Center. After removing bad slides, including duplicate, non-retrievable, or invalid slides, and slides with inconclusive reports, etc., the dataset is split into training, validation, and testing sets. Samples in the testing set are independent of the ones in the training and validation sets. For their deep learning model, U-Net is used as the primary model for training on entire slides. There are six classes. One of them is negative and the remaining five classes are the five Gleason grade groups. After the training step is done, they describe a generative adversarial network for normalizing external testing sets called CycleGANs. The model achieves strong performance based on the results from Table II, shown below.

TABLE II. Internal test set results from [2]. Only AUC is shown here, and more details are included in [2].

Internal test set	Number of cases	AUC
Benign v. malignant	250/285	0.990
Benign + GG1 v. GG ≥ 2	325/210	0.978
Benign + GG1 + GG2 ≥ GG3	377/158	0.974

However, the classes are very unbalanced and this may lead to high AUC values since potentially the model can be biased toward one side and still achieve high performance. On the other hand, their model is able to be directly trained on entire slides instead of small patches and saves time and effort for patch extraction, but it also relies on their CycleGANs to perform well on different kinds of datasets.

Recently, an outstanding training algorithm on slide-wise classification greatly reduces training size by using a streaming implementation of convolutional layers. This is proposed by Pinckaers et al. [10]. They implement two training procedures. One utilizes streaming convolutional operations to reduce data size by allowing slide-wise training. The other one works like a traditional patch-wise classification and generates overall slide-wise results.

The streaming implementation first splits a whole slide image into tiles and uses convolutional layers to reduce tile sizes, until they are able to merge back to the original slide and start training. Pinckaers et al. [10] were able to achieve good performance from both training methods and make slide-wise training possible for future research.

In Tolkach et al. [14], another powerful DL models is developed based on NASNetLarge, a state-of-the-art convolutional neural network. The authors train using 389 whole slide images with pixel-wise annotation, which is done by 3 experienced pathologists. There are two models developed, one for benign v. tumor and the other for Gleason group classification. The model is trained using semi-supervised method. It is first trained on slides that have the same primary and secondary patterns. Then it is being used to annotate slides that have different primary and secondary patterns. The annotated slides are later used for fine tuning. This method is able to achieve a high accuracy of 98%.

C. *Clinical Use Cases*

Recent works have done studies on letting pathologists and AI models work together. Concordance between pathologists and subspecialists have increased in almost all cases with help from AI models. In Nagpal et al. [11], a deep learning system is proposed to mimic pathologists' workflow by analyzing small patches and generating an overall Gleason grade for each slide. There are 19 pathologists and 6 subspecialists involved in the concordance measurements. For more detailed results refer to [11]. Summary results are shown in Table III below. Further research is needed to explore potential uses of deep learning in clinical workflows with pathologists and subspecialists.

TABLE III. SUMARY OF CONCORDANCE BETWEEN A DEEP LEARNING SYSTEM AND PATHOLOGISTS, AND BETWEEN A DEEP LEARNING SYSTEM AND SUBSPECIALISTS. DATA IS FROM [11].

	Pathologists	**Subspecialists**
Grading tumor-containing biopsy specimens (n = 498)	58.0%	71.7%
Non-tumor v. tumor (n = 752)	94.7%	94.3%

In Pantanowitz et al. [12], similar studies have done using a deep learning system as a tool for assisting in real-life clinical use cases. Once the AI model is trained, testing slides are used for inference and the results are compared with pathologists' diagnoses. If there is any disagreement, pathologists are asked to re-diagnose the slide. The agreement between the algorithm and pathologists is around 0.882 on cancer percentage, which is the cancer proportion of a whole slide image. This study reports successful collaboration between pathologists and AI models for clinical practice and shows a need for further studies on the screening of prostate cancer slides and standardizing reports in patient management.

Another study done by Steiner et al. reports improvements on agreement between AI-assisted general pathologists and subspecialists [13]. The deep learning model being used in this study is described in [11]. The main purpose is to find out whether this kind of pre-studied deep learning model is useful in clinical use cases. A total of 240 biopsies or slides are reviewed by 20 pathologists.

In the five cases studied above, only one case indicated that AI's assistance is unnecessary. Overall, a 5.6% increase between unassisted pathologists and assisted pathologists is seen. This demonstrates AI's benefit in quality and consistency of prostate cancer detection and grading. There are also some limitations in this study: the Gleason grade group grading does not have further clinical information regarding each individual patient; only 1 biopsy is used per case in this study and more biopsies for each case are necessary to check the consistency of AI assistance.

III. PATCH-LEVEL CLASSIFICATION METHODOLOGY

A. *Data Collection*

The public dataset of H&E-stained prostate biopsies from Radboud University Medical Center is used for training, validation, and testing sets in our study. To conduct patch-wise classification, each patch (a subsection of a whole slide image) needs a label, e.g. benign, pattern 3, pattern 4, etc. Only the dataset from Radboud University Medical Center has regional annotations on them. An example is shown in Fig. 2. The dataset from Radboud is labeled by trained students and contains significant errors. Students and pathology experts were asked to annotate a smaller dataset. The concordance between the students' results and experts' results is approximately 72%. This highlights significant disagreement between the labels of the students and experts.

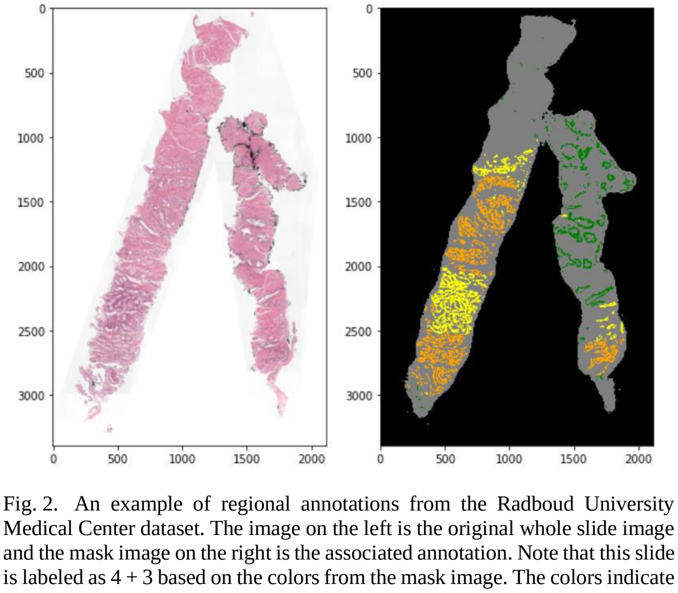


Fig. 2. An example of regional annotations from the Radboud University Medical Center dataset. The image on the left is the original whole slide image and the mask image on the right is the associated annotation. Note that this slide is labeled as 4 + 3 based on the colors from the mask image. The colors indicate the different tissue types listed in Table IV.

TABLE IV. PROSTATE TISSUE TYPES REPRESENTED BY DIFFERENT COLORS BASED ON THE RADBOUD UNIVERSITY MEDICAL CENTER DATASET [3].

Background or unknown	⬛
Stroma (non-epithelium tissue)	▒
Benign epithelium	🟩
Patten 3 epithelium	🟨
Patten 4 epithelium	🟧
Pattern 5 epithelium	🟥

B. *Data Cleaning and Patch Extraction*

In the training set, there are some items that have inconsistent labels and annotations. Some of the slides even have artifacts including pen marks, tiling effects from the stitching, missed regions during scanning, etc. These slides are removed from the training set, so that the model will not learn from these bad or confusing samples.

Patch extraction can be split into a few small steps, including splitting slides into tiles, sorting tiles based on the amount of stroma or epithelium, labeling tiles based on the annotations, and eventually saving patches from the same patient together. Each patch is 128 x 128 pixels.

Table V shows the total number of whole slide images being split into patches and the number of patches for each class. Note that originally there are 5,158 slides and 949 are filtered out due to inconsistencies between labels and annotations. Note that the number of patches for each class is not the same, meaning that the data is initially unbalanced. The dataset can be balanced by using data augmentation, which will be discussed in the next section.

TABLE V. NUMBER OF SLIDES FOR EACH STEP.

Total slides	5158					
Filtering slides	**Consistent**					**Inconsistent**
	4209					949
Patches extracted for each class	Stroma	Benign	Pattern 3	Pattern 4	Pattern 5	
	118951	4769	13160	23181	3238	

C. *Data Augmentation*

Random augmentation methods are applied for each generated patch, including rotation, brightness, hue, saturation, and flip horizontally and vertically. The goal is to generate more patches based on the original patches to balance the number of patches of each class. Each method also can be adjusted randomly. For example, changing rotation degree, brightness magnitude, saturation magnitude can be random, etc.

D. *Train/Test/Validation Sets*

For the training sets, the same data size is applied to all classification models. However, different classification models have different data sizes for their testing set and validation set. This is because data augmentation cannot be applied to the testing set and the validation set, since both sets must remain independent of the training set.

Patches from the same patient can only be used in one of training, testing, and validation sets. This eliminates the possibility that the training set and testing set contain different patches that are from the same slides. Table VI shows the number of slides, stroma patches, benign patches, pattern 3, 4, and 5 patches for the training, testing, and validation sets.

E. *Model Training*

Table VII shows all the models that are trained in this work for Gleason pattern classification and how data samples are distributed for each class. The first four rows indicate the models used for a decision flow method [15]. The last row indicates the model for a normal multi-class classification model [15]. Note that data augmentation has been applied to overcome the unbalanced training set.

DenseNet [16] is used as the primary model in this work after investigating other modern deep neural networks. It was chosen because it does not easily overfit and its validation loss was the lowest compared to other network architectures. A Dropout layer is also added on top of the DenseNet 201 to address overfitting. The original DenseNet 201 does not have any regularizations like L2, L1, or Dropout layers.

Many experimental iterations are needed to tune hyperparameters in order that the model learns smoothly and quickly and also achieves high performance. Table VIII lists the hyperparameters for all models that were determined via our testing.

TABLE VI. NUMBER OF SLIDES AND PATCHES FOR THE TRAINING, TESTING, AND VALIDATION SETS USED IN THIS WORK.

	Train (56%)	Test (30%)	Validation (14%)
Slides	2358	1262	589
Stroma patches	15120	8100	3780
benign patches	2670	1430	667
pattern 3 patches	7396	3948	1842
pattern 4 patches	12981	6954	3254
pattern 5 patches	1813	986	453

TABLE VII. CLASSIFICATION MODELS TRAINED IN THIS WORK. NOTE THAT ALL MODELS USE AT LEASE 10000 PATCHES FOR TRAINING.

Classification	Data size (train only)
Stroma v. Benign + 3 + 4 + 5	10000 v. 2500 + 2500 + 2500 + 2500
Benign v. 3 + 4 + 5	10000 v. 3334 + 3334 + 3334
3 v. 4 + 5	10000 v. 5000 + 5000
4 v. 5	10000 v. 10000
Stroma v. Benign v. 3 v. 4 v. 5	10000 v. 10000 v. 10000 v. 10000 v. 10000

TABLE VIII. LIST OF HYPERPARAMETERS USED IN THIS WORK FOR ALL MODELS.

Hyperparameters	**Value**
Dropout layer	0.5
L2 regularization	0.0001
Batch size	32
Learning rate	$0.0002 * (\frac{1}{1 + 0.02 * epoch})$ $Initial\ rate = 0.0002$ $Decay\ rate = 0.02$
Momentum	0.7

The loss function used is categorical cross entropy and the optimizer is stochastic gradient descent with momentum, since it does not have a large memory requirement. Early stopping is also applied to prevent models from overfitting. Once each patch is classified from the models, it is colored based on its classification result.

IV. CLASSIFICATION MAPS GENERATION

A few steps are needed to generate a classification map. A whole slide image must be split into patches in order to be fed into trained models. Stride size or overlapping area can be decided manually. This affects how detailed or clear the generated classification map appears, basically setting the resolution of the classification map.

In Fig. 3, gray squares indicate areas for patch-level classification, which are 128 x 128 pixels in size. To obtain higher resolution, patches extracted from each slide for classification map generation should overlap each other and only a small number of pixels of each patch are colored instead of the entire patch.

In Fig. 4, from left to right, the first image is the original whole slide image, the second image is the provided mask image, the third image is the predicted mask image using the decision flow method, and the last one is also a predicted mask image, but using the multi-class classification method instead. In this particular image, each color patch is 16 x 16 pixels. The smaller the size, the better the resolution of the predicted mask images are created. It takes 16 times longer to generate a "16 x 16" image than to generate a "128 x 128" image.

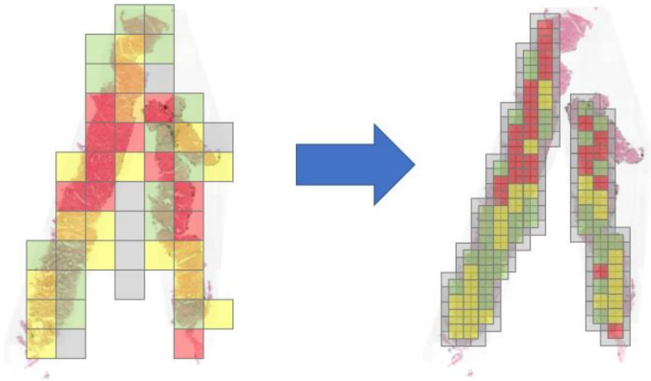


Fig. 3. An example of how to generate a better mask image through a slide window method that allows overlapping areas between patches. The right image has better visualization but also requires more computation power.

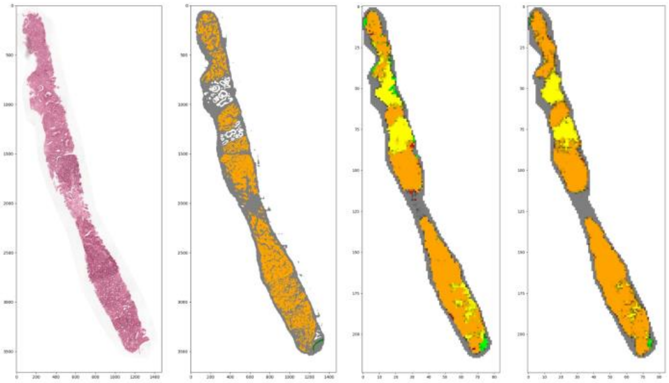


Fig. 4. A comparison between the original whole slide image, provided mask image, and two predicted mask images. From left to right, first is the original whole slide image, second is the original mask image, third is the predicted mask image using the decision flow method, and the fourth is the predicted mask image using the multi-class classification method.

V. RESULTS

A. Patch-Level Results

There are a total of five models trained in this work. Four for the decision flow method, one for the multi-class method. Fig. 5 shows the training process of all five models. In each figure, the y axis represents cross-entropy loss and the x axis represents the number of epochs. The blue line indicates training loss and the orange line indicates validation loss. None of them indicates a strong overfitting since validation loss never goes back up.

Table IX shows the AUC of each of the five models tested on the testing sets. The first four rows are for the decision flow method and the last row is for the multi-class method, which is also the overall AUC.

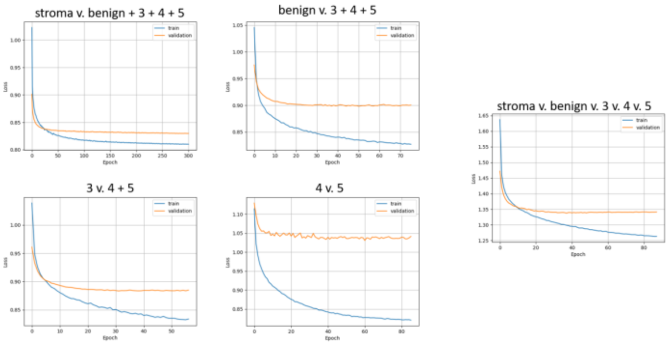


Fig. 5. Training process outputs of the five models. In each subplot, the blue line represents training loss and the orange line represents validation loss. The four models on the left are from the decision flow method and the one model on the right is a normal multi-class method.

TABLE IX. AUC VALUES OF THE FIVE MODELS TESTED ON THE TESTING SETS. PLEASE NOTE THAT THESE VALUES ARE NOT COMPARABLE SINCE EACH MODEL SERVES DIFFERENT PURPOSES AND WAS TESTED ON DIFFERENT TEST SETS.

Model	**AUC**
stroma v. benign + 3 + 4 + 5	0.9954
benign v. 3 + 4 + 5	0.9719
3 v. 4 + 5	0.9876
4 v. 5	0.8742
stroma v. benign v. 3 v. 4 v. 5	0.9630

Fig. 6 demonstrates the power of data augmentation by presenting the AUC's of most of the models before and after data augmentation is applied. Four models are shown here and three of them are used in inference. From left to right, each model has its own color and the lighter color is, the larger the data augmentation size is added. Almost all models' performance increase by some amount after data augmentation is applied.

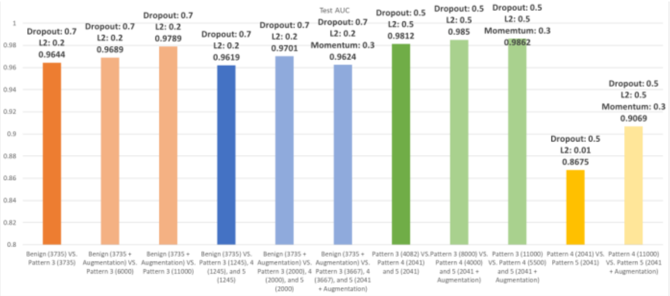
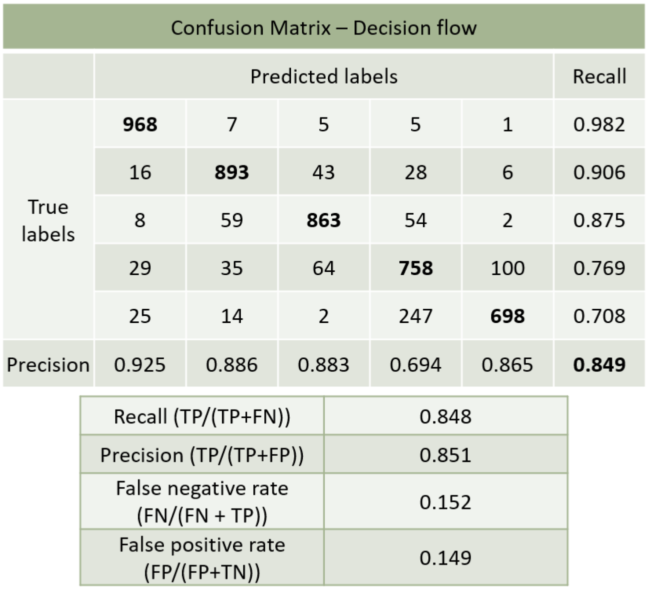


Fig. 6. Comparison of four models' AUCs before and after data augmentation is applied. Each model has its own color and the lighter it is, the bigger data augmentaion size is used.

The trained models more poorly distinguish between pattern 4 and pattern 5, as compared to other cases in the decision flow method, as seen by the AUC values. Pattern 4 and pattern 5 cancers receive the same clinical treatment even though they are in different classes, so it makes sense that these two classes are the hardest to classify.

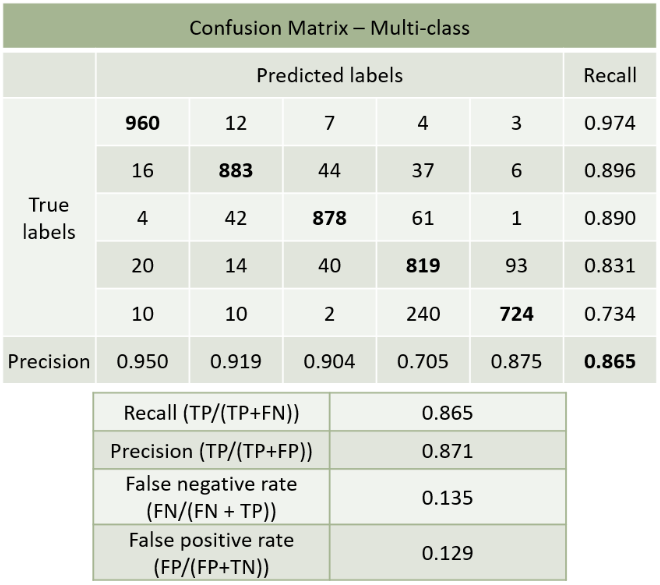


Confusion Matrix – Decision flow

	Predicted labels					Recall
True labels	**968**	7	5	5	1	0.982
	16	**893**	43	28	6	0.906
	8	59	**863**	54	2	0.875
	29	35	64	**758**	100	0.769
	25	14	2	247	**698**	0.708
Precision	0.925	0.886	0.883	0.694	0.865	**0.849**

Recall (TP/(TP+FN))	0.848
Precision (TP/(TP+FP))	0.851
False negative rate (FN/(FN + TP))	0.152
False positive rate (FP/(FP+TN))	0.149

Fig. 7. The confusion matrix shown here is for the decision flow method. It is tested on an unseen test set. The accuracy is 0.849. Recall, false negative rate, and false positive rate are calculated by averaging those of the five classes.



Confusion Matrix – Multi-class

	Predicted labels					Recall
True labels	**960**	12	7	4	3	0.974
	16	**883**	44	37	6	0.896
	4	42	**878**	61	1	0.890
	20	14	40	**819**	93	0.831
	10	10	2	240	**724**	0.734
Precision	0.950	0.919	0.904	0.705	0.875	**0.865**

Recall (TP/(TP+FN))	0.865
Precision (TP/(TP+FP))	0.871
False negative rate (FN/(FN + TP))	0.135
False positive rate (FP/(FP+TN))	0.129

Fig. 8. The confusion matrix shown here is for the multi-class method. It is tested on an unseen test set. The accuracy is 0.865. Recall, false negative rate, and false positive rate are calculated by averaging those of the five classes.

Fig. 7 and Fig. 8 show the confusion matrices for the patch-level results of the decision flow method and the multi-class method. Both are tested on the same unseen test sets. The overall performance between these two methods is very similar. However, decision flow shows slightly higher accuracy on benign and stroma cases and multi-class method performs better on pattern 3, 4, and 5 cases. The multi-class method is more frequently used, since it is a more modern and efficient way of solving multi-class problems using a softmax layer at the end of networks. The reason the decision flow method is used here is to provide a finer control on each sub model during inference. Parameters like probability threshold can be changed based on certain needs.

B. Classification Map Results

An experienced pathologist examined a few of the generated classification maps and provided feedback on the original tissue classifications, as well as on the predicted mask images. This section presents a few typical scenarios, including both concordance and non-concordance cases that help draw some conclusions on how well this algorithm performs.

Fig. 9. shows the first case, which has unannotated areas left in white color. Yellow circles are drawn by the pathologist around these unannotated areas and the color chosen by the pathologist matches the predicted classification maps. In this case, the predicted mask images provide better annotations in the circled areas. Also note that the provided label is 4 + 4, but both the pathologist and predicted classification maps indicate that this slide should be 4 + 3.

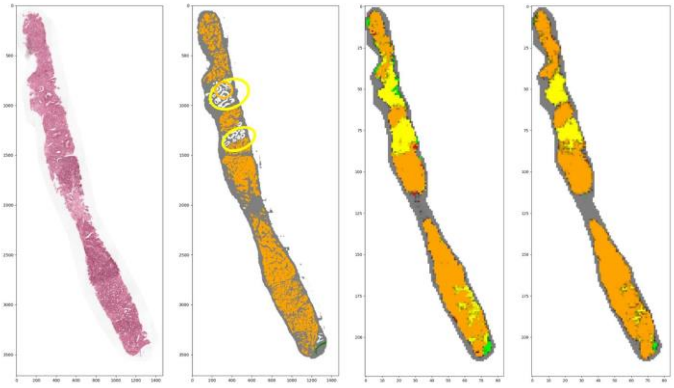


Fig. 9. The first comparsion between the provided mask image and the predicted classification maps. The provided label is 4 + 4.

Fig. 10. shows the second example. Both the provided mask image and the predicted classification map show concordance with each other. However, the pathologist believes the middle-circled region should be completely orange (pattern 4). For the bottom region, the pathologist believes it should be 90% orange and 10% yellow (pattern 3), so the slide label should be 4 + 4, not 4 + 3, since the yellow part exists in only 10% of the slide.

Fig. 11. shows the third example. The provided mask image has an error at the bottom area where the classification map shows green (benign) instead of red. The pathologist agrees with the classification maps and believes that the top area should be 90% orange and 10% red. The last thing to note, the provided label 5 + 4 does not agree with the pathologist's assessment of the sample, having labeled it 4 + 4. The pathologist thinks pattern 5 only makes up 10% of the entire slide.

Fig. 12. shows the fourth example. In this case, the pathologist completely agrees with the provided label and the provided mask image. Based on his view, this is a rare case called duct type that occurs infrequently. The predicted classification maps show green, orange, and red colors. This would probably lead to a slide label classification of 4 + 5. 4 + 4 and 4 + 5 are in the same prognostic grade group, meaning their slide-level predictions are still the same.

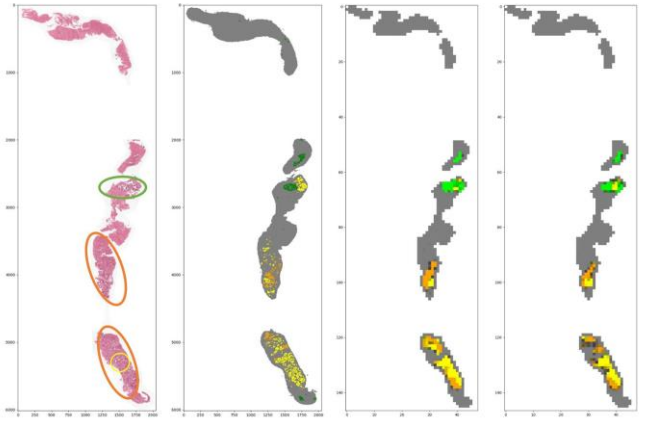


Fig. 10. The second comparsion between the provided mask image and the predicted classification maps. The provided label is 4 + 3.

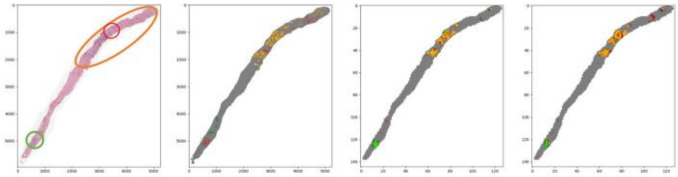


Fig. 11. The third comparsion between the provided mask image and the predicted classification maps. The provided label is 5 + 4.

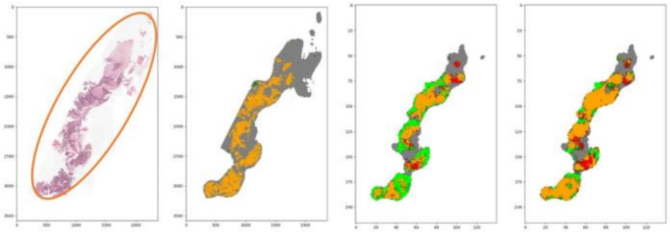


Fig. 12. The fourth comparsion between the provided mask image and the predicted classification maps. The provided label is 4 + 4. Note that this is a rare case called duct type that does not have many similar cases in the dataset.

Fig. 13. shows the fifth example. The provided mask image shows a high degree of cancer severity in the slide. However, the pathologist thinks the entire slide is completely benign and believes that the provided label and mask image are both wrong. For contrast, the predicted classification maps accurately annotate most regions in green color, indicating benign tissue with a few areas shown in orange (pattern 4).

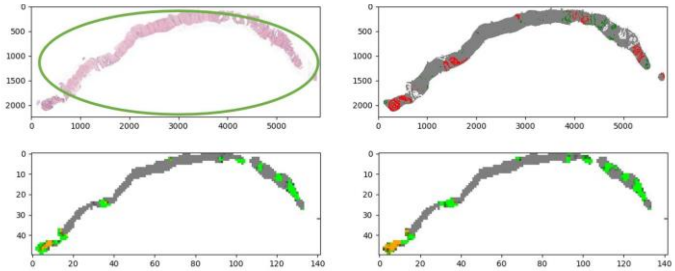


Fig. 13. The fifth comparsion between the provided mask image and the predicted classification maps. The provided label is 5 + 5.

VI. Discussion

Based on the comparisons from the last section, classification maps sometimes label samples incorrectly and provide false information to pathologists. Two potential reasons exist for this situation. The first reason comes from the source errors found in the Radboud dataset based by Bulten et al. [3]. The students first read the original pathology reports and then color the biopsy image. Before getting into patch-level annotations, errors already exist in the slide-level labels. As mentioned in [3], both students and the three experts annotated a smaller portion of the same dataset. The quadratic weighted kappa value and accuracy between the students' and the experts' reviews are 0.853 and 0.720.

The second reason is due to the way patches are extracted before starting training patch-level models. These models show high accuracy and AUC vales as mentioned in Section V. Each patch contains only one color at a time, but during classification maps generation, there are many patches containing more than one color. Although single-color patches do exist often in real cases, the classification result on those multi-color patches can only provide one class at a time rather than a mixture of classes.

Some strong results are produced in this work that can contribute to other efforts. Classification map generation is what differentiates this work from others. Almost all other work using this dataset attempt to train a model that directly classifies slide-level labels. It is known that an entire whole slide image cannot be fed into any neural network due to its size. Instead, most work focuses on how to resize the original slides by using sub-images that contain biopsies only. Although this kind of work is able to classify slide-level labels faster, it is more of a black box relative to our approach. Our work generates classification maps for pathologists to review and is easy to debug and adjust on a patch level. Because of this advantage, it is easier to determine which system component between the patch-level and classification maps generation needs improvements. Classification maps generation is somewhat limited to the performance of the patch-level classification. Based on the classification map results in Section V, increasing the variety of patches for each class should produce improvements. Also, utilizing larger patch sizes to capture more information for each class would be a good approach to improving overall performance.

VII. Conclusion

In this work, a three-stage deep learning system is designed and implemented to better assist pathologists in clinical practice. Patch-level models largely affect classification maps results. On the other hand, classification maps can at least serve as a reference for pathologists in real cases. It provides pathologists with a better understanding of how the system works and where it fails. Knowing the capability of this system, pathologists can use its assistance without being sidetracked if an error has been made by the system.

References

[1] Kryvenko, O. N., & Epstein, J. I. (2016). Prostate cancer grading: a decade after the 2005 modified Gleason grading system. Archives of pathology & laboratory medicine, 140(10), 1140-1152.

[2] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., ... & Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology, 21(2), 233-241.

[3] Bulten, Litjens, Pinckaers, Ström, Eklund, Kartasalo, et al. (2020, March). The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system. Presented at the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2020), Lima, Peru: Zenodo.

[4] Kayser, K., Görtler, J., Metze, K., Goldmann, T., Vollmer, E., Mireskandari, M., ... & Kayser, G. (2008, December). How to measure image quality in tissue-based diagnosis (diagnostic surgical pathology). In Diagnostic pathology (Vol. 3, No. 1, pp. 1-7). BioMed Central.

[5] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., ... & Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging, 35(8), 1962-1971.

[6] Tani, S., Fukunaga, Y., Shimizu, S., Fukunishi, M., Ishii, K., & Tamiya, K. (2012). Color standardization method and system for whole slide imaging based on spectral sensing. Analytical Cellular Pathology, 35(2), 107-115.

[7] Zarella, M. D., Yeoh, C., Breen, D. E., & Garcia, F. U. (2017). An alternative reference space for H&E color normalization. PloS one, 12(3), e0174489.

[8] Shaban, M. T., Baur, C., Navab, N., & Albarqouni, S. (2019, April). Staingan: Stain style transfer for digital histological images. In 2019 IEEE 16th international symposium on biomedical imaging (Isbi 2019) (pp. 953-956). IEEE.

[9] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis, 58, 101544.

[10] Pinckaers, H., Bulten, W., van der Laak, J., & Litjens, G. (2020). Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. arXiv preprint arXiv:2006.03394.

[11] Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P. H. C., Steiner, D. F., ... & Mermel, C. H. (2020). Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. JAMA oncology, 6(9), 1372-1380.

[12] Mahal, B. A., Muralidhar, V., Chen, Y. W., Choueiri, T. K., Hoffman, K. E., Hu, J. C., ... & Nguyen, P. L. (2016). Gleason score 5+ 3= 8 prostate cancer: much more like Gleason score 9?. BJU international, 118(1), 95-101.

[13] Steiner, David F., et al. "Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies." JAMA network open 3.11 (2020): e2023267-e2023267.

[14] Tolkach, Yuri, et al. "High-accuracy prostate cancer pathology using deep learning." Nature Machine Intelligence 2.7 (2020): 411-418.

[15] Tan, W. (2021) Detection of Prostate Cancer in Patch-Level Gleason Grading using Deep Learning. MS Thesis, Drexel University.

[16] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.