2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA) | 978-1-6654-0969-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/AICCSA53542.2021.9686869

Automatic Prostate Cancer Grading Using Deep Architectures

Muhammad Mohsin, Arslan Shaukat, Usman Akram, Muhammad Kaab Zarrar
National University of Sciences and Technology (NUST), Islamabad, Pakistan
mohsinuet123@gmail.com, arslanshaukat@ceme.nust.edu.pk

***Abstract*— Prostate cancer is the second most aggressive type of cancer among men aged over 45, and it has a major effect on people's lives. Early diagnosis and grading of prostate cancer from tissue images is necessary. Large scale inter observer reproducibility exists in grading the prostate biopsies. This leads us to move towards a computer based model that can accurately detect and grade the cancerous prostate from non-cancerous one. The paper is focused on deep learning based models to automatically grade the prostate cancer from tissue microarray images. Deep learning models directly learn the features via convolutional layers. Two datasets have been used for implementation of our proposed model, Harvard dataset and Gleason Challenge 2019. Our proposed UNET based architecture is used for training as well as validation and testing. We used four different deep learning models, VGG19, ResNet50, Mobilenetv2 and ResNext50 for our UNET based encoder. With our proposed framework, we have achieved 0.728 and 0.732 average Cohen's kappa with F1 on both datasets respectively. The results show that our proposed UNET based deep learning model shows better performance as compared to other state of the art models.**

Keywords*— Tissue Microarray (TMA), Gleason Score, Convolutional Neural Network (CNN), Prostate Cancer (PCa), Deep Learning.

I. Introduction

In the United States, prostate cancer is the second most aggressive form of cancer found in men [1]. In developed countries, prostate cancer is increasing exponentially due to high living standards and population explosion. Whereas in Pakistan prostate cancer ratio is 10.7% which is increasing in recent years. Large number of prostate cancer patients die every year due to insufficient diagnosis environment and large-scale inter-observer variation between pathologists. This leads to design a model that can early detect and correctly classify the cancerous grade. Since 1960, the Gleason grading algorithm has been the most reliable and effective prognostic predictor for prostate cancerous cells [2]. The World Health Organization (WHO) strongly recognizes the Gleason scoring algorithm, which was modified and revised by the International Society of Urological Pathology in 2005 and 2014 (ISUP) [3]. Despite numerous advances in the clinical diagnosis of prostate cancerous cells, histology-based Gleason scoring remains the most effective prognostic indicator of prostate carcinoma early detection and grading. [4]. The architectural pattern of cancerous cells is used to render the histological evaluation. The architecture pattern consists of well-differentiated cell to the poor ones. This architectural pattern is entirely responsible for the Gleason score. Fig. 1 shows the tissue micro arrays (TMAs) of benign, Gleason score 6, Gleason 7 and Gleason score 8. The Gleason score is assigned on the basis of tissue structural patterns. Gleason 6 contains well-formed glands, Gleason score 7 contains well-formed glands with lesser component of cribriform glands and Gleason score 8 contains only poorly form cribriform glands with lesser component of well-formed glands.

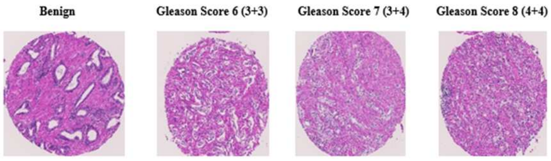

Fig. 1. Example of Tissue Microarray Images with Benign to Gleason Score 6, Gleason Score 7 and Gleason Score 8 [5].

The pathological analysis determines the Gleason ranking, which is a time-consuming and important process with a lot of inter-and intra-observer variance. This problem occurs when Gleason grade 3 is distinguished from Gleason grade 4 with a Gleason score of 7 (4+3 or 3+4), which can have a complicated effect on subsequent care. Since major medical decisions are focused on the evaluation of biopsy specimens, there is a clear need for an automated prostate cancer grading model [5]. Many feature engineering based techniques are used for automatic Gleason grading [6]. The success of feature engineering based techniques is totally dependent on how accurately features are extracted and its compatibility with model. Recently, computer-assisted method of grading using convolutional neural network has considered to play a significant role in medical image analysis [7, 8]. This method has replaced the traditional way of drawing out features for image categorization with totally different approach of allowing the network to finalize which features have to be considered. The outstanding results on standard dataset have made CNN a widely used technique for pattern identification. Here, we have worked on CNN based system to analyze different tissue microarray images and assign them Gleason score.

This paper is focused on deep learning model for automatic grading of prostate cancerous cells. It consists of six sections. Section-II discusses the literature review both on feature engineering and deep learning based models. Section-III briefly presents the available working datasets of prostate biopsies. Section-IV explains the proposed experimental methodology based on deep learning. Section-V reports our experimental results. Last section draws the conclusions.

II. Literature Work

A. Feature Engineering Based Methods

In recent studies, papers have been published on developing of automatic computer based Gleason Grading methods to correctly classify the prostate cancer. A very

978-1-6654-0969-8/21/$31.00 ©2021 IEEE

common technique is to extract the tissue features and applied feature engineering based classifiers on these features to classify the tissue as benign or cancerous. Smith et al [9] extracted features based on 2D-Discrete Wavelet Packet Decomposition and then passed to Support Vector Machines (SVM) classifier to predict the grade of prostate. Farooq et al [6] used Gabor filter and local binary patterns for features extraction. These selected features are used with KNN classifier to grade the prostate cancerous cells. The power distribution of histological tissue images was used by Smith et al. [9] to reflect texture characteristics of prostatic biopsies. They used nearest neighbor classifier to grade those characteristic features into Gleason grade 1, grade 3, grade 4 and grade 5. For automatic Gleason grading, Farjam et al. [10] suggested a multistage classifier based on morphometric and texture characteristics. Those morphometric and texture features are used to identify gland units. The image is then classified into grades 1 through 5 using morphometric and texture attributes derived from gland units in a sequence of classification levels. Nguyen et al. [11] used structural features of prostate glands to classify pre-extracted regions of interest (ROIs) into benign, G3, and G4. The described papers achieved good results on their datasets due to high dependence on feature extraction.

B. *Deep Learning Based Methods*

Extensive research has been carried out based on deep learning methods to design automatic computer based models for accurately grading the cancer [12]. Deep learning based fully convolution neural network models are very useful in early detection of prostate cancer as compared to feature engineering based models [13]. Deep learning models are also very successful in prostatic segmentation [14]. In early stages, CNN based architectures [15] were used for better feature extraction as compared to conventional feature engineering based methods. They analyzed deep entropy features using different CNN models and passed those features to random forest classifier to predict the Gleason score. Augmented based technique is proposed in [16], which uses three different CNNs, combined their prediction results and predict the Gleason score by logistic regression method. They achieved 92% and 86% accuracy in classifying low and high prostate grade. [17] used both morphological and texture features achieving 79% accuracy of classifying benign with other higher grades. A very recent study [2] classified the prostate TMAs into four categories benign, grade 3, grade 4 and grade 5. The Cohen's kappa achieved with two pathologists was reported as 0.72.

In this research, we have proposed deep learning model based on UNET with two phase training to achieve pathologist level results. We used four different CNN architectures Vgg19, ResNet50, mobileNetv2 and ResNext50 for feature extraction. Compare to other deep learning based models, our main contributions are:

- Deep learning based models require very high amount of data for training, so due to limited data, we have proposed data augmentation with multi batches to extract the contextual information from each TMA.
- For solving the class imbalance problem that exists in both datasets, we have proposed two phase training with equal and true class ratio to achieve state of the art results. We first trained our model on actual class, and after that use those trained features to train the model again on equal class ratio.
- We have tested our model generalization on another TMAs based dataset, which again gives state of art performance as compared to previous work done on this dataset.

III. Datasets

In this section, we briefly discuss the prostatic datasets containing tissue microarrays (TMAs), that are used in our study.

A. *Gleason Challenge 2019 MICCAI Dataset*

The recently published dataset from the Gleason 2019 challenge has been used in our work [5]. Tissue microarray (TMA) images are included in this competition. The dataset contains TMAs with their corresponding masks in PNG format. Each mask contains pixels which shows score. Several specialist pathologists with years of experience in their fields annotate each TMA picture in great detail. Data is prepared by pathologists on the basis of majority voting annotations. The data contains samples belonging to benign and 3 different grades, G3, G4 and G5. The distribution of dataset into training, validation and testing cohorts is given in Table I.

TABLE I. The distribution of Gleason grade in the training, validation and test cohorts.

	Total Cases	Benign	G 3	G 4	G 5
Train	188	72	111	134	10
Validation	33	13	20	23	1
Test	23	15	10	14	3
Total	244	100	141	171	14

Training images contain 188 TMA of prostatic tissues with benign, Gleason 3, Gleason 4, and Gleason 5. For testing the deep learning model, we used 23 TMAs and for validation, we used 33 TMAs. Grade 3 and Grade 4 have high class ratio of images in data which create over fitting problems which are addressed by batch normalization, dropout and data augmentation.

TABLE II. The distribution of Gleason grade in the training, validatio and test cohorts.

	Benign	G 6	G 7	G 8	G 9	G 10	Total Images
TMA 76	42	35	25	15	2	14	133
TMA 80	12	88	38	91	3	13	245
TMA 111	0	95	39	69	16	8	227
TMA 204	0	1	17	25	8	69	105
TMA 199	61	69	2	26	2	1	176

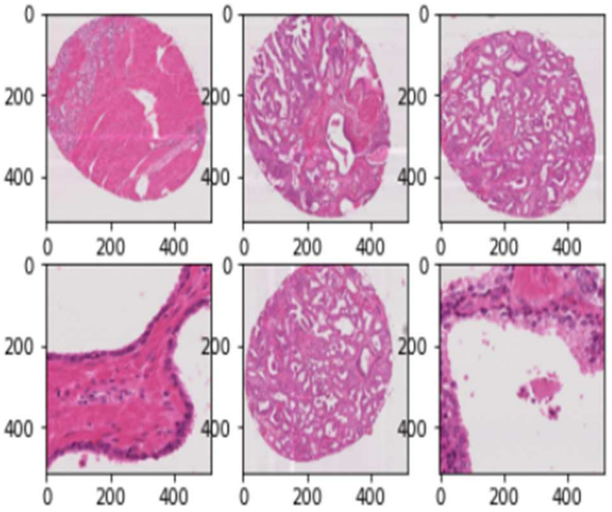

Fig. 2. Example of TMAs of MICCAI Dataset

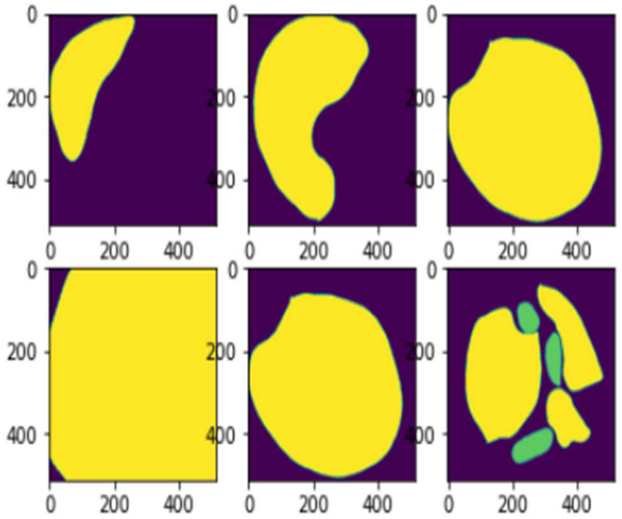

Fig. 3. Example of true masks of corresponding TMAs.

Fig. 2 and Fig. 3 represent the TMAs and their corresponding masks present in MICCAI dataset [5]. The TMAs are achieved after preprocessing of original images with corresponding masks after data augmentation. Fig. 3 shows ground truth of original TMAs which consist of Gleason score ranging from 0 to7.

B. Harvard Dataverse V1 Dataset

Harvard dataset is acquired from online databases of Harvard [2]. It contains five tissue micro arrays TMA, each with 200 to 300 spots, which make up the data. Objects or non-prostate tissue with spots (e.g. lymph node metastasis) is excluded from the study. The first pathologist (K.S.F.) identified the prostate TMA spots by carefully delineating cancerous areas and giving each one a Gleason pattern of 3, 4 or 5. TMA spots without cancerous areas have also been discovered to be benign. The distribution of Gleason scores across different TMAs is seen in Table II.

Since TMA 80 has the most events, it has been assigned as the study cohort. TMA 76 was used as a confirmation cohort because it has the most evenly distributed Gleason ratings. As a result three other TMAs are used as a training cohort. A second pathologist annotated the TMA spots in the research data separately, allowing the inter-pathologist variability to be quantified.

The TMAs and their corresponding masks of Harvard Dataverse V1 Dataset are represented in Figure 4 and Figure 5.[2]. The TMAs mentioned above are obtained after preprocessing of the original images with corresponding masks after data augmentation. Figure 3 illustrates the ground truth masks of original TMAs with Gleason scores ranging from 0 to 5.

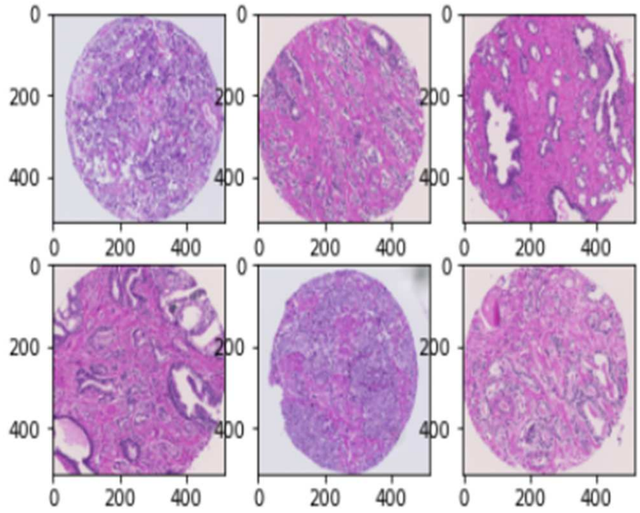

Fig. 4. Example of TMAs of Harvard Dataverse V1 Dataset

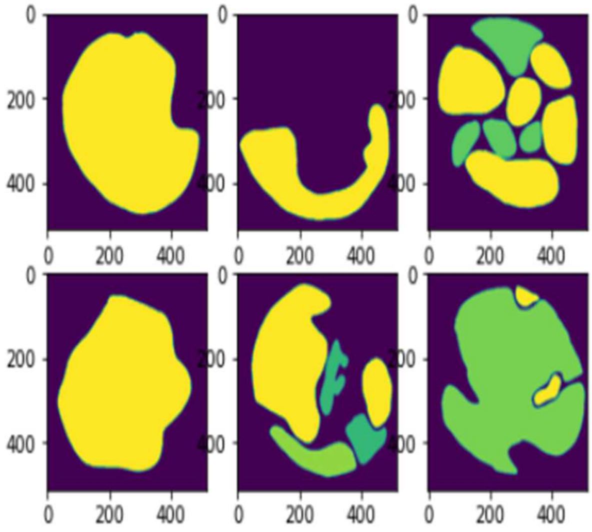

Fig. 5. Example of True masks of corresponding TMAs

IV. Experimental Methodology

We briefly describe our proposed model in this section. The block diagram of our proposed scheme is shown in Fig. 6.

A. *Data Preprocessing and Normalization*

We first pre-process the data before passing it to UNET model. All the images are resized to achieve better generalization. One hot encoding is used for better performance, because it allows the categorical data to be more expressive. TMA images and their corresponding masks of Gleason Challenge dataset have resolution of 4608 x 5120. The TMA images have been resized to 512 x 512 for the 2 stages of training process. In PNG masks, the pixel values 0, 1, 2, 3, 4, 5 and 6 show corresponding Gleason score.

In Harvard Dataverse VI, the size of TMAs and masks are different, 3100 x 3100. We have resized the images to 512 x 512 for better generalization of model results which is trained on same size of images. In PNG masks, the pixel values 1, 2, 3 and 4 show corresponding Gleason score. After getting tissue microarray TMA images, data augmentation is applied to increase the data size.

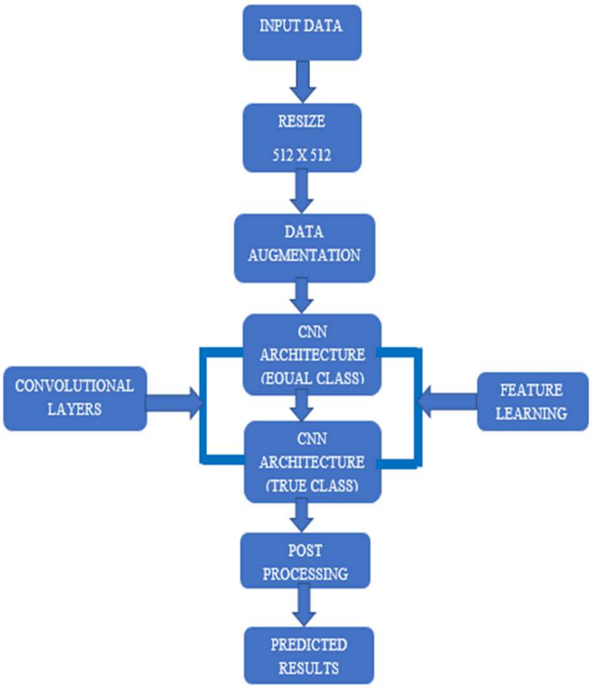

Fig. 6. Proposed Model Diagram

B. *Data Augmentation*

Data augmentation is considered important, particularly in medical application fields where the amount of data is small. The overall performance of deep learning algorithm is improved using data augmentation. Data augmentation is the process of creating new training data based on existing training data by using methods like horizontal, vertical shift, horizontal flip, vertical flip, rotation, scaling and zooming. For data augmentation, we have used augmentor library [18], where tissue micro array images (TMA) are rotated from 90 to 270 degree with 0.25 probability and also performed left to right flipping. Similarly top to bottom flipping is done with 0.15 probability. Finally we have used random cropping operation with 0.35 probability. During data augmentation, we select equal class ratio for augmentation to combat the data limited problems. Table III shows the parameters for data augmentation technique that we have used in this paper.

TABLE III. Summary of Data Augmentation Implementation

Methods	Range
Rotation	90 , 270 Degree
Flip Left to Right	15 % Probability
Flip Top to Bottom	25 % Probability
Random Cropping	40 % Probability

C. *Proposed Convolutional Neural Network (CNN)*

Deep learning based models require large amount of data for their training to perform well on test data. So data augmentation is applied using Augmenter library to reduce the overfitting problem and increase the generalizability of model. After properly resizing the TMA images and data augmentation, TMAs are fed into CNN. Next step is to extract feature maps which is very critical for pixel level classification, because bad features may led to poor pixel level classification results. CNN is used for direct feature learning from data. In available datasets, large amount of class imbalance exists due to which model shows over fitting. To reduce class unequal problem, we used two phase training which solves the class imbalance problem. In first phase of training, we use equal class ratio to train the model and then used those weights to train on true class ratio. This methodology leads to achieve good results.

We have used four different CNN architectures with UNET. The four architectures VGG19, ResNet50, Mobilenetv2 and ResNext50 are used for extracting progressive features from pre-processed TMAs. UNET model used these progressive features, up sample them, concatenate the features and generates the predicted mask. The mask contains the pixel level classes. Due to ResNet50 residual property and greater number of trainable parameters on both MICCAI and Harvard Dataverse V1 datasets, we have achieved state of the art results. The details of architectures' parameters are given in Table IV and Table V.

TABLE IV. Detail of parameters of four Encoder architecure with UNET on MICAAI Dataset

Model	Trainable	Non Trainable	Total
VGG19	29,058,807	4,032	29,062,839
ResNext50	31,993,850	70,214	32,064,064
MobileNetv2	8,012,215	36,096	8,048,311
ResNet50	32,514,426	47,558	32,561,984

TABLE V. Detail of parameters of four Encoder architecure with UNET on harvard Dataset

Model	Trainable	Non Trainable	Total
VGG19	9,033,988	20,028,416	29,062,404
ResNext50	25,121,345	7,140,105	32,261,450
MobileNetv2	5,822,020	2,225,856	8,047,876
ResNet50	9,059,079	23,502,470	32,561,549

UNET architecture was first introduced by [13]for biomedical image segmentation. This model make its place in the field of medical image segmentation in recent times due to its uniqueness. UNET model has the capability of contracting the input images into multiple feature maps. After contraction, it uses its previous feature maps to expand till it reaches its output level. Due to its contraction and expansion capability with concatenation power, it preserves the structural integrity of images. The layer details of UNET model with all four CNN architectures are given in Table VI.

TABLE VI. UNET layers of Encoding of our four architecures

Model	Convolution layers	Pooling Layers
VGG19	16	5 max pool
ResNext50	48	1 max pool, 1 global avg pool
MobileNetv2	10	1 avg pool
ResNet50	48	1 max pool, 1 avg pool

D. Results Evaluation

To assess the efficiency of segmentation models, different evaluation criteria are used. We have used standard evaluators that are currently being used in automatic Gleason grading of TMA images and also being used in clinical process of calculating inter observer variation. Our literature reveals that TMA based datasets are evaluated on Cohen's Kappa [19] , F1 Score and Dice score [20]. We have also used them to assess the performance of our proposed model. Another reason for choosing these evaluators is to perform comparison with previous reported results in literature on automatic Gleason grading.

$$\text{Cohens Kappa} = \frac{\text{Po} - \text{Pe}}{1 - \text{Pe}} \qquad (1)$$

Equation (1) shows the Cohen's Kappa, where Po is observed agreement among ratters and Pe is hypothetical probability of chance agreement.

$$\text{F1 Score} = 2\left[\frac{\text{Precision . Recall}}{\text{Precision} + \text{Recall}}\right] \qquad (2)$$

F1 score is the function of Precision and Recall. The calculation of precision and recall is dependent on true positive and the sum of true positive and false positive which is shown in Equation (3) and (4).

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \qquad (3)$$

$$\text{Dice Score} = \frac{2 \times \text{TP}}{(\text{TP} + \text{FP}) + (\text{TP} + \text{FN})} \qquad (4)$$

$$\text{Overall Score} = \frac{\text{Cohens Kappa} + \text{F1 Score}}{2} \qquad (5)$$

V. Experimental Results

In this section we briefly discuss the experimental results. We have performed our experimentation on Gleason Challenge 2019 and Harvard datasets. Both datasets contain TMAs which have pixel level annotations by expert pathologists. Gleason Challenge 2019 dataset contains 244 TMAs, 188 are used as training by applying data augmentation, 33 TMAs are used as independent validation and 23 are used as testing cohort. While on Harvard dataset, we have 509 TMAs as training, 133 TMAs as validation set and 245 TMAs as independent test cohort. Both datasets contain Gleason score of 3, 4 and 5.

A. Results on Gleason MICAAI Dataset

In MICAAI dataset, four different CNN architectures are used as our UNET model encoder. We have achieved high scores on all four encoder architectures evaluated using Dice Score, Cohen's kappa and F score. ResNet50 has performed well as compared to other encoder architectures due to its residual property and faster convergence. Categorical cross entropy is used as loss function with learning rate of 0.0001. ResNet50 has achieved overall score of 0.728, which is highest as compared to other three architectures.

TABLE VII. Comparison of UNET Based Model Results on MACCAI Dataset

UNET Model	Encoder Backbone	Dice Score	Cohen's Kappa	F Score	Overall Score
1	VGG-19	0.49	0.30	0.31	0.31
2	ResNext 50	0.47	0.28	0.29	0.29
3	MobileNET	0.63	0.63	0.61	0.62
4	**ResNet 50**	**0.68**	**0.72**	**0.73**	**0.72**

Table VII results clearly indicate that when ResNet 50 is used as backbone with UNET, it outperforms all other architectures. The Dice Score for ResNet50 exceeds MobileNetv2 by 7.5%. While the increase is more significant when compared with ResNext50 and VGG-19 scores. This score go past ResNext50 and VGG-19 scores by 31% and 28% respectively. Similarly, ResNet 50 shows better results than other frameworks for both Cohen's Kappa and F1 score. A significant rise of 0.42 and 0.44 can be seen in overall score of our encoder from VGG-19 and ResNext50. However, there is an increase of 14.7% from MobileNetv2, proving it to be the best encoder for Gleason score assignment. Due to its residual blocks and identity mapping, ResNet50 has produced optimal feature maps. Those optimal feature maps contain all the pertinent features which can perfectly classify the image to its ground-truth class that is why ResNet50 gives state of art results as compared to other encoder architectures.For evaluating our model. Overall score is calculated which is the average of both Cohen's kappa and F1 score.

TABLE VIII. COMPARISON OF OUR MODEL WITH BEST PERFORMING MODEL IN TERM OF COHEN'S KAPPA ON MACCAI 2019 DATASET.

Model Team	**F1 Score**	**Cohen's Kappa**	**Score**
YujinHu	0.84	0.84	0.84
Nitinsinghal	0.79	0.79	0.79
Ternaus	0.78	0.78	0.78
Zhangjingmri	0.77	0.77	0.77
sdsy888	0.75	0.75	0.75
cvblab	0.75	0.75	0.75
XiaHua	0.71	0.71	0.71
AlirezaFatemi	0.71	0.71	0.71
Jpviguerasguillen	0.64	0.64	0.64
qq604395564	0.64	0.64	0.64
Unipabs	0.58	0.58	0.58
Our Proposed	**0.72**	**0.73**	**0.73**

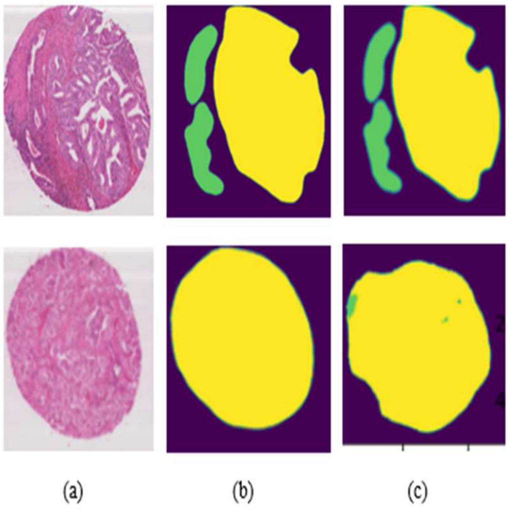

Fig. 7. Results on best performing model UNET-ResNet50 (a) Original TMA images (b) Original masks (c) Predicted mask

Fig 7.(a) shows original images of MICCAI dataset, (b) shows the original masks which contains pixels ranging from 0 to 7 and (c) shows the predicted masks based on our best performing architecture. Fig. 8 and Fig. 9 show the training and testing accuracy of ResNet50 model with their corresponding losses. We trained our model on 50 epochs, the graphs show that training as well as testing accuracy gradually increases as number of epochs increases. The training and testing loss gradually decreases as the number of epochs increase due to increment of learning.

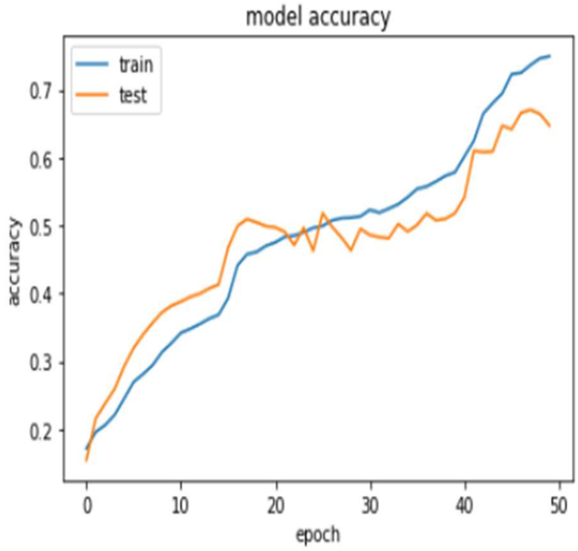

Fig. 8. Graphical Representaion of ResNet50 accuracy on MICAAI dataset

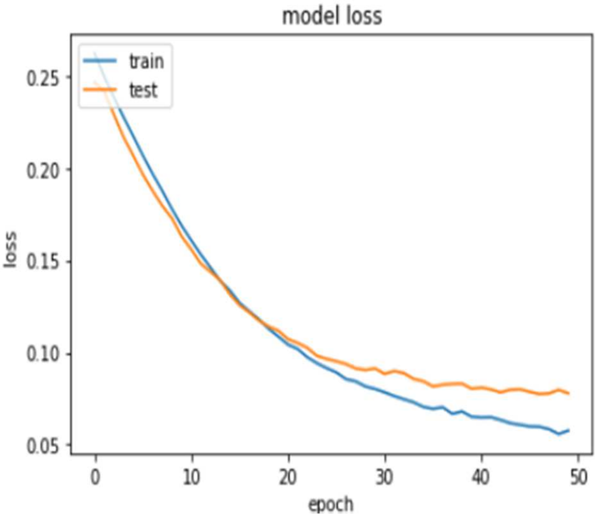

Fig. 9. Graphical Representaion of ResNet50 loss on MICAAI dataset

In Gleason Challenge 2019 [5], teams from all over the world have participated and are working on MICCAI dataset. Table VIII shows the comparison of top selected teams taken from Gleason Challenge with our top performing model UNET-ResNet50. Most of the teams have managed to achieve good results. However their methodology and results have not been published so far. We have still managed to achieve competitive results on the Gleason Challenge dataset.

B. Results on Harvard Dataset

Table IX shows the results of Harvard dataset. On Harvard dataset, we have implemented VGG19, ResNext50, MobileNetV2 and ResNet50 architectures as encoder with UNET. It is readily apparent from Table IX that ResNet50 is again the top model in comparison to other encoders with regard to all variables. Dice score of ResNet 50 is greater by 2.6%, while Cohen's score by 5.5% from VGG-19, hence attaining the highest accuracy among the four models. ResNet 50 stands out from the rest in terms of overall score as well, with 5.5% increase from VGG-19 and Mobile Net and 7% greater than ResNext50.

TABLE IX. COMPARISON OF OUR MODEL WITH BEST PERFORMING MODEL IN TERM OF COHEN'S KAPPA ON HARVARD DATAVERSE DATASET

UNET Model	Encoder Backbone	Dice Score	Cohen's Kappa	Overall Score
1	VGG-19	0.75	0.69	0.69
2	ResNext 50	0.62	0.68	0.68
3	MobileNETV2	0.70	0.63	0.69
4	**ResNet 50**	**0.77**	**0.73**	**0.73**

Fig 10.(a) shows the example images from the dataset (b) shows the ground truth masks of Harvard dataset which contains pixels ranges from 0 to 4 and (c) shows the predicted masks using our best performing architecture.

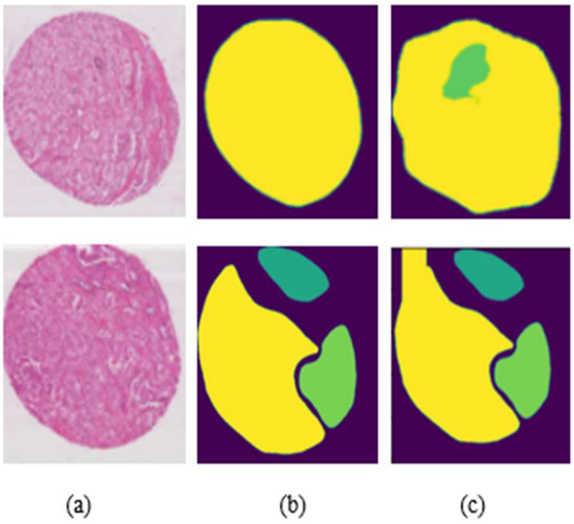

Fig. 10. Results on best performing model UNET-ResNet50
(a) Original TMA images (b) Original masks (b) Predicted masks

ResNet50 contains 48 convolution layers stacked one after the other, with max and average pooling. It is a long deep trained model with residual block, which gives us state of the art results as compared to other models. Due to identity mapping property, gradient loss problem is solved and network learning speed increases. That is why we have achieved best results with UNET-ResNet50 architecture. In testing our model generalization, we have used the same Harvard dataset as used in [2] and have managed to achieve slightly better results as compared to them.

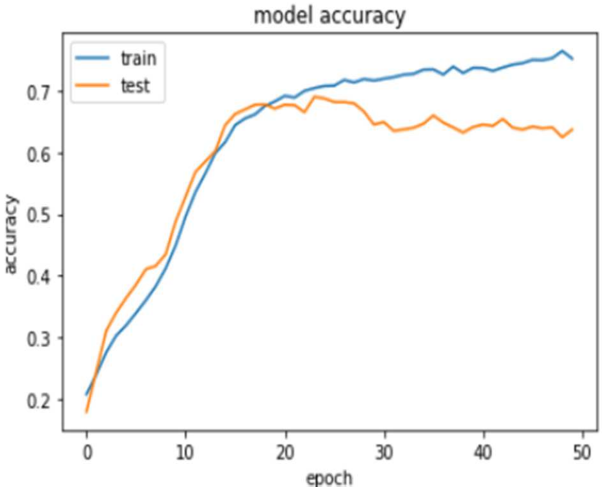

Fig. 11. Graphical Representaion of ResNet50 accuracy on Harvard dataset

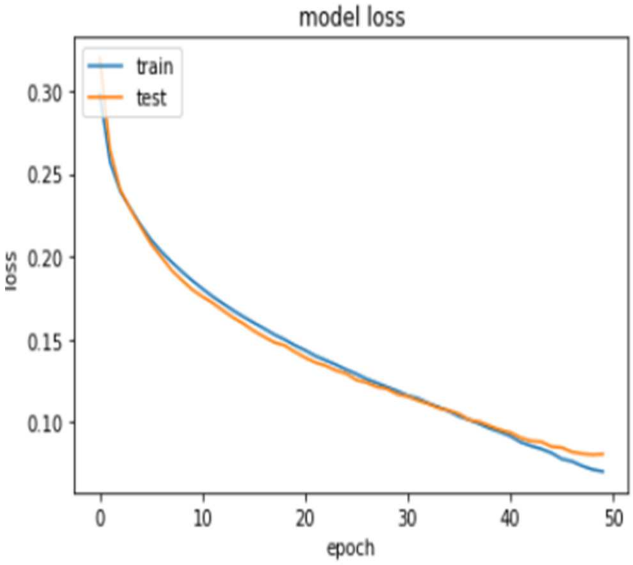

Fig. 12. Graphical Representaion of ResNet50 accuracy on Harvard dataset

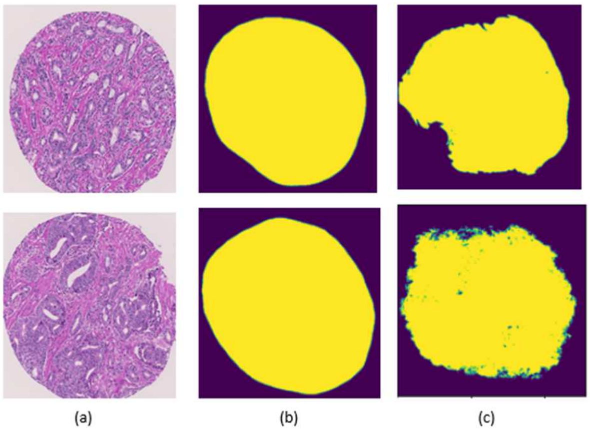

Fig. 13. Results on least performing model UNET-ResNext 50
(a) Original TMA images (b) Original masks (b) Predicted masks

Fig. 11 and 12 show the training and testing accuracy with corresponding loss. As the number of training epochs are increased, the training accuracy is increased and loss decreases. We test our model using same parameters and achieved state of the art results. This shows that our proposed architecture is effective as it outperforms previous results on the same dataset. The detailed comparison of our results with other results reported in literature is shown in Table X. Fig 13. (a), (b), (c) shows the least performing model ResNext-50 on both datasets.

TABLE X. COMPARISON OF UNET BASED MODEL RESULTS ON HARVARD DATASET

Model	No. of Images	Cohen's Score
Eirini et al. [2]	640	0.72
UNET-ResNet 50	**640**	**0.73**

VI. CONCLUSIONS

In this paper, we have implemented deep learning based models on two different datasets for automatic prostate cancer grading. We have proposed a methodology which is based on UNET model for automatic prostate cancer grading at pixel level and predicted the pathologist level results on both datasets. We used four different CNN architectures, VGG19, ResNext50, MobileNetV2 and ResNet50 as an encoder to UNET model. UNET with ResNet50 encoder gives us state of the art results as compared to other encoder architectures due to its unique identity mapping. Due to lesser number of samples, we have also implemented data augmentation on both datasets which increases the overall performance of UNET model. Our experimental results show that our proposed deep learning based model achieved competitive results on Gleason Challenge dataset and higher results on Harvard dataset, as compared to previous reported results. In future, availability of large amount of data with less class imbalance problem may improve segmentation and prove to be helpful in the development of clinically acceptable methods for grading of prostate cancer. Moreover depending on resources, large scale CNN based architectures can be deployed for more robustness and better generalization.

REFERENCES

[1] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, B. S. Knudsen, *et al.*, "Path R-CNN for prostate cancer diagnosis and gleason grading of histological images," *IEEE transactions on medical imaging,* vol. 38, pp. 945-954, 2018.

[2] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, *et al.*, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific reports,* vol. 8, pp. 1-11, 2018.

[3] S. F. Faraj, S. M. Bezerra, K. Yousefi, H. Fedor, S. Glavaris, M. Han, *et al.*, "Clinical validation of the 2005 ISUP Gleason grading system in a cohort of intermediate and high risk men undergoing radical prostatectomy," *PloS one,* vol. 11, 2016.

[4] J. Gordetsky and J. Epstein, "Grading of prostatic adenocarcinoma: current state and prognostic implications," *Diagnostic pathology,* vol. 11, p. 25, 2016.

[5] M. G. C. f. Pathology. (17 October 2019). *Gleason 2019 Challenge.*Available:https://gleason2019.grandchallenge.org/

[6] M. T. Farooq, A. Shaukat, U. Akram, O. Waqas, and M. Ahmad, "Automatic gleason grading of prostate cancer using Gabor filter and local binary patterns," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, 2017, pp. 642-645.

[7] Q. Zhu, B. Du, and P. Yan, "Boundary-weighted domain adaptive neural network for prostate MR image segmentation," *IEEE transactions on medical imaging,* vol. 39, pp. 753-763, 2019.

[8] Y. Wang, B. Zheng, D. Gao, and J. Wang, "Fully convolutional neural networks for prostate cancer detection using multi-parametric magnetic resonance images: an initial investigation," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3814-3819.

[9] Y. Smith, G. Zajicek, M. Werman, G. Pizov, and Y. Sherman, "Similarity measurement method for the classification of architecturally differentiated images," *Computers and Biomedical Research,* vol. 32, pp. 1-12, 1999.

[10] R. Farjam, H. Soltanian-Zadeh, R. A. Zoroofi, and K. Jafari-Khouzani, "Tree-structured grading of pathological images of prostate," in *Medical Imaging 2005: Image Processing*, 2005, pp. 840-851.

[11] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognition Letters,* vol. 33, pp. 951-961, 2012.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, pp. 436-444, 2015.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241.

[14] Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan, "Deeply-supervised CNN for prostate segmentation," in *2017 international joint conference on neural networks (IJCNN)*, 2017, pp. 178-184.

[15] A. Chaddad, M. J. Kucharczyk, C. Desrosiers, I. P. Okuwobi, Y. Katib, M. Zhang, *et al.*, "Deep radiomic analysis to predict gleason score in prostate cancer," *IEEE Access,* vol. 8, pp. 167767-167778, 2020.

[16] D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean, "Deep Learning-Based Gleason grading of prostate cancer from histopathology Images—Role of multiscale decision aggregation and data augmentation," *IEEE journal of biomedical and health informatics,* vol. 24, pp. 1413-1426, 2019.

[17] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi, and P. W. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human pathology,* vol. 35, pp. 1121-1131, 2004.

[18] U. Ali, A. Shaukat, M. Hussain, J. Ali, K. Khan, M. Khan, *et al.*, "Automatic cancerous tissue classification using discrete wavelet transformation and support vector machine," *J. Basic. Appl. Sci. Res,* vol. 6, pp. 15-23, 2016.

[19] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica,* vol. 22, pp. 276-282, 2012.

[20] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports," *Academic radiology,* vol. 11, pp. 178-189, 2004.