Generalizing and automating the classification of prostate cancer severity

Andrea Perera-Ortega[1]

[1]Medical Informatics Laboratory, School of Computing, Queen's University, Kingston, Canada

ABSTRACT

PURPOSE: Correctly classifying the severity of prostate cancer is fundamental to the prognosis and treatment of this disease, but these important diagnoses often suffer significantly from inter and intra-observer variability. Recent automated deep learning methods have shown to be comparable in performance to pathologists when predicting prostate cancer severity, but have not been tested with multi-centre data sets. This study aims to create an automated deep learning method that can generalize well to multiple data centres. **METHODS:** The data consisted of 10,616 prostate biopsies from two data centres. Four different Convolutional Neural Network (CNN) architectures were utilized to find an optimal model for this application including: ResNet50, EfficientNetB0, EfficientNetB3, and EfficientNetB6. **RESULTS:** The best model performance achieved was using an EfficientNetB6 architecture resulting in an accuracy of 67%, QWK of 77% and AUC of 90%. The model was evaluated on an independent test set of 900 prostate biopsy samples. **CONCLUSION:** The test results achieved show promising results as a foundation to further develop this model with multiple data centers.

Keywords: Prostate Cancer, Deep Learning, ISUP Grade, Digital Pathology, Convolutional Neural Networks

1. INTRODUCTION

Each year, prostate cancer (PCa) accounts for over 350,000 deaths.[1] In Canada, prostate cancer is the most commonly diagnosed cancer in males, with 1 in 9 expected to be diagnosed with prostate cancer in their lifetime.[2] Essential to decreasing mortality is the development of more precise diagnostic measures for this disease. Diagnosis of PCa is based on the Gleason grading of tissue biopsies, which is based on the architectural growth of the tissue and is determined by pathologists. After assigning the biopsy with a Gleason grade, it is then converted to an ISUP (International Society of Urological Pathologists) grade. This score determines the prognosis of the cancer, and is critical to the decisions made regarding a patient's treatment. However, there is a chance that the severity of the prostate cancer may be over or under-graded, as well as inter and intra-observer variability between pathologists. This variability can lead to unnecessary treatment or worse, missing a severe diagnosis.

Automated deep learning systems have shown to be promising in accurately grading PCa, achieving performance levels on par with pathologists, but have not been evaluated in multi-centre data set. Ström *et al.*[3] trained deep neural networks to to assess prostate biopsies by predicting the presence, extent and Gleason grade of each. They achieved an average pairwise kappa score of 0.62 when evaluating on an independent test set of 87 biopsies which is comparable to the performance of a trained pathologist. Additionally, this model predicted Gleason grades as opposed to ISUP scores and were only tested on one data center. Bulten *et al.*[4] also used deep learning methods to predict Gleason grades of prostate biopsies and achieved a quadratic Cohen's kappa score of 0.918 on an independent test set of 535 biopsies, outperforming 10 of 15 pathologist observers. While both of these studies show promising results, neither were tested on multiple data centers at scale.

This project aims to further extend previous work by using machine learning techniques to build a model that will predict ISUP grades of prostate biopsies, generalized to three data centres, including one local to Queen's University. This approach will consist of training four different Convolutional Neural Network (CNN) architectures including: ResNet50, EfficientNetB0, EfficientNetB3, and EfficientNetB6 to find one best suited for this application.

2. MATERIALS AND METHODS

2.1 Data and Image Acquisition

This problem was initially presented as a Kaggle challenge, titled the Prostate cANcer graDe Assessment (PANDA) Challenge[5] in conjunction with MICCAI 2020. Data in the form of prostate biopsies was provided by the Radboud University Medical Center and Karolinska Institute for use during this challenge. 5,414 prostate biopsies were provided by the Karolinska Institute and 5,202 were provided by the Radboud University Medical Center for a total of 10,616 prostate biopsies. 189 biopsies were excluded from the study as they were deemed "suspicious" due to having an incorrect ISUP grade or being a blank image. This resulted in a total of 10,427 biopsies that were used in the study. The distribution of ISUP grades across the samples is shown in Figure 1.

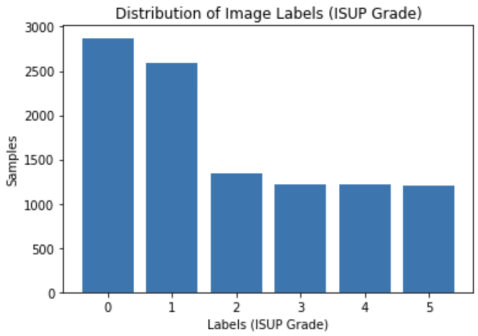

Figure 1. Distribution of ISUP labels across all samples in the data set

2.2 Data Pre-Processing

The prostate biopsies were provided as multi-level TIFF (Tag Image File Format) images at 1x, 4x and and 16x resolutions. In order to balance the trade-off between image resolution and model computation time, the images at the 4x resolution level were used and downsampled by another factor of 2x for a total downsampling factor of 8x. This allowed for faster computations without sacrificing important image details necessary for model training. Additionally, all images were normalized so that their pixel values were between 0 and 1 and standardized to have a mean of zero and unit variance.

2.3 Concatenated Tile Extraction

Due to the whole slide images (WSI) being very large in nature and individual images differing in sizes, it was not feasible to use these images directly when training a model. Both the Karolinska Institute and Radboud University provided masks with pixel-level annotations, but the Karolinska masks did not provide specific information about ISUP grades, only specifying benign versus cancerous regions. For this reason, individual patches could not be used to train a network either. To address these issues, sixteen 64x64 tiles were extracted from each of the samples and concatenated back together to form a 256x256 image composed of tiles. The tiles were strategically chosen from each biopsy to minimize the amount of background in the tile and maximize the amount of prostate tissue. Figure 2 shows an example prostate biopsy before and after undergoing the aforementioned concatenated tile extraction process.

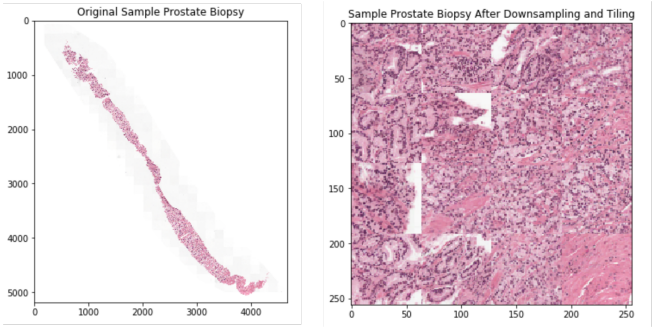

Figure 2. This figure shows an example prostate biopsy used in the study before and after downsampling and concatenated tile extraction

2.4 Data Augmentation

Due to the unbalanced nature of the image classes, it was essential that data augmentation would be performed as a way to not only increase the size of the training set, but also balance the classes. In order to address this problem, images in each class were augmented until there were 3000 samples per class, resulting in a total of 18,000 perfectly balanced samples. The augmentations performed on each image were selected randomly from a set of predefined augmentations. These possible augmentations included: rotations of 90°, 180°, and 270°, as well as horizontal and vertical flips.

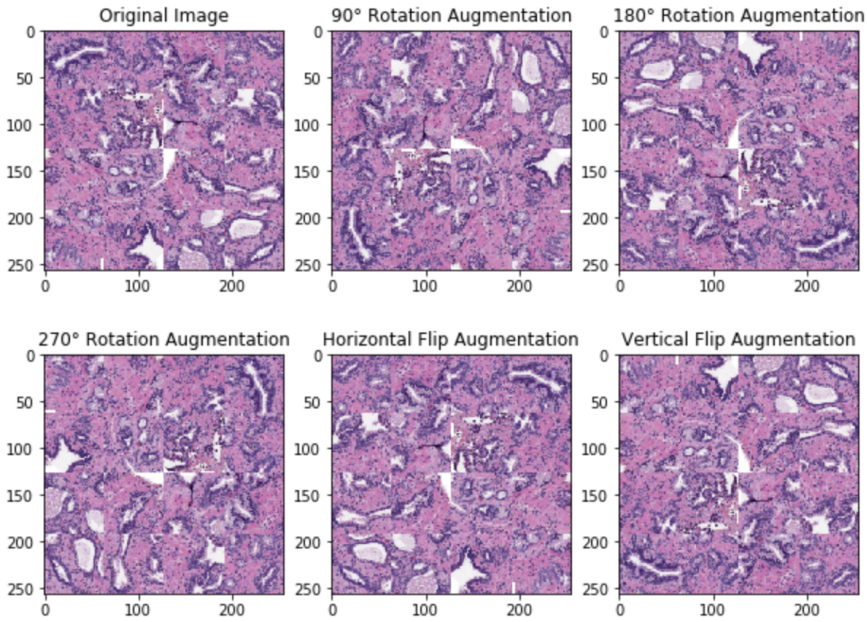

Figure 3. All of the potential augmentations that could be performed on a sample biopsy during the data augmentation step

2.5 Data Division

After augmentation was performed, the data was split into training, validation and test sets using a 80/15/5 split, respectively. 14400 samples were used in the train set and 2700 samples were used in the validation set. The independent test set not seen during training was composed of 900 samples. The class balance was maintained in all sets so that the classifier would be exposed to all classes equally during training and testing.

2.6 Model Architecture

Several popular Convolutional Neural Network (CNN) architectures were implemented during the course of this research. These include: ResNet50, EfficientNetB0, EfficientNetB3, and EfficientNetB6. The architectures were implemented using Keras[6] and TensorFlow.[7] For each architecture, the final dense layers were removed in order to define the model's input shapes and number of classes. Additionally, all of the architectures used were pre-trained with ImageNet[8] weights and used Dropout and Batch Normalization layers. The architectures of each model are denoted in Tables 1 through 4. For all four models, categorical crossentropy loss was used, and the optimizer used was Adam. The TensorFlow function "ReduceLROnPlateau" was used to decrease the learning rate when the validation loss stayed stagnant for more than five epochs.

Table 1. ResNet50 Architecture

Layer Type	**Output Shape**	**Parameters**
ResNet50 (model)	(None, 8, 8, 2048)	23587712
GlobalAvgPool2D	(None, 2048)	0
Dropout (0.5)	(None, 1536)	0
Dense	(None, 256)	524544
Dropout (0.5)	(None, 256)	0
Dense	(None, 6)	1542
		Total: 24,113,798

Table 2. EfficientNetB0 Architecture

Layer Type	**Output Shape**	**Parameters**
EfficientNetB0 (model)	(None, 8, 8, 1280)	4049564
GlobalAvgPool2D	(None, 1280)	0
Flatten	(None, 1280)	0
BatchNormalization	(None, 1280)	5120
Dropout (0.25)	(None, 1280)	0
Dense	(None, 256)	327936
BatchNormalization (	None, 256)	1024
Dropout (0.25)	(None, 256)	0
Dense	(None, 6)	1542
		Total: 4,385,186

Table 3. EfficientNetB3 Architecture

Layer Type	**Output Shape**	**Parameters**
EfficientNetB3 (model)	(None, 8, 8, 1536)	10783528
GlobalAvgPool2D	(None, 1536)	0
Flatten	(None, 1536)	0
BatchNormalization	(None, 1536)	6144
Dropout (0.25)	(None, 1536)	0
Dense	(None, 256)	393472
BatchNormalization	(None, 256)	1024
Dropout (0.25)	(None, 256)	0
Dense	(None, 6)	1542
		Total: 11,185,710

Table 4. EfficientNetB6 Architecture

Layer Type	Output Shape	Parameters
EfficientNetB6 (model)	(None, 2304)	40960136
Flatten	(None, 2304)	0
BatchNormalization	(None, 2304)	9216
Dropout (0.25)	(None, 2304)	0
Dense	(None, 256)	590080
BatchNormalization	(None, 256)	1024
Dropout (0.25)	(None, 256)	0
Dense	(None, 6)	1542
		Total: 41,561,998

2.7 Evaluation

Each architecture was trained on an independent test set containing 900 samples, with 150 samples per class. The test set contained original biopsies and augmentated biopsies. The metrics used in the evaluation process were categorical accuracy, Quadratic Weighted Kappa (QWK), and the Area under the ROC Curve (AUC).

3. RESULTS AND DISCUSSION

After employing the four different architectures, it was found that the three EfficientNet architectures performed relatively well and similarly to each other. EfficientNetB6 overall performed best, but not by a large margin. The main difference between the three EfficientNet variants is the number of parameters and the duration of training time. As the number of parameters increased, the training times increased, with minimal improvements in accuracy. Because of this, EfficientNetB0 may be suitable enough for this application. Out of the four architectures, the ResNet50 model performed the worst. All of the architecture metrics can be found in Table 5. The EfficientNetB6 performance curves for categorical accuracy, QWK and AUC can be found in Figure 4.

Table 5. Architecture Performances for ISUP Grade Classification

Architecture	**Metric**	Train	Validation	Test
ResNet50	Accuracy	0.9986	0.5574	0.5556
	QWK	0.9993	0.7284	0.7126
	AUC	1.000	0.8212	0.8214
EfficientNetB0	Accuracy	0.9340	0.6078	0.6122
	QWK	0.9561	0.7622	0.7790
	AUC	0.9955	0.8843	0.8799
EfficientNetB3	Accuracy	0.9399	0.6367	0.6511
	QWK	0.9626	0.7599	0.7705
	AUC	0.9967	0.8947	0.8965
EfficientNetB6	Accuracy	0.9615	0.6474	0.6656
	QWK	0.9757	0.7994	0.7739
	AUC	0.9982	0.8965	0.8965

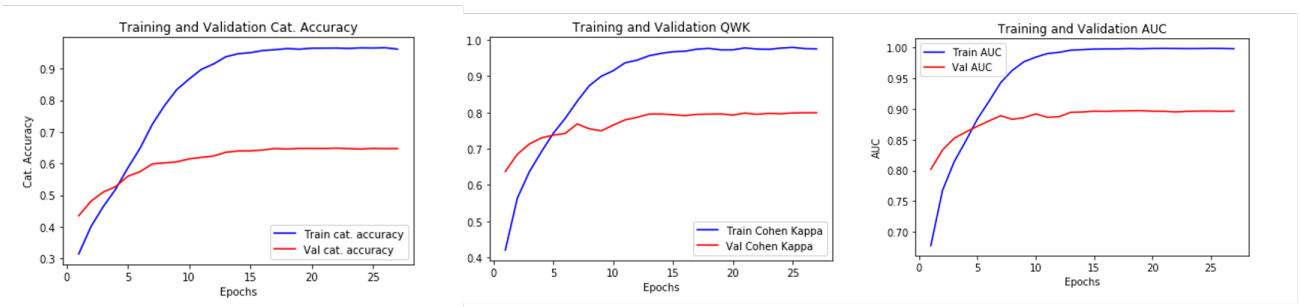

Figure 4. The resulting categorical accuracy, QWK, and AUC for the EfficientNetB6 model

Observing the Quadratic Weighted Kappa (QWK), the EfficientNet architectures performed well. This is a popular metric used in multi-class classification due to the fact that it takes into account the amount of similarity between predictions and actual labels. It is also an appropriate metric for this study's application due to the nature of ISUP grade diagnosis. As the ISUP grade increases, the severity of the prostate cancer increases. Additionally, ISUP grades that are closer together look more similar, and are most often confused with each other. For example, it can be seen in the confusion matrix in Figure 5 that the majority of the falsely predicted labels are very close to the true positive diagonal. This implies that when the model predicts the ISUP grade incorrectly, often times it is only off by a factor of one, so the false predictions aren't extremely dissimilar from the true label. Upon observation of the confusion matrix, one way to improve the results may be to implement label binning as opposed to converting integer labels to categorical data. For example, since labels closer to each other express more similarity, using a label of [1, 1, 1, 0, 0] for an ISUP grade of 3 versus [0, 0, 0, 1, 0, 0] may assist with the classification.

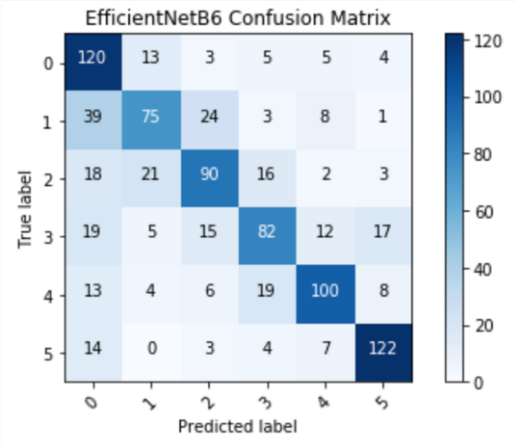

Figure 5. Confusion matrix for the EfficientNetB6 model predictions

4. CONCLUSION

These results show a positive start towards building a generalizable model that can predict ISUP grade across multiple data centers. The best overall model performance used an EfficientNetB6 architecture resulting in 67% accuracy, 77% Quadratic Weighted Kappa, and 90% AUC on an independent test set of 900 prostate biopsies.

5. FUTURE WORK

Future work for this project will include further refining the model with the existing Karolinska and Radboud data to achieve higher performing metrics. Possible refinements include more sophisticated data augmentation (for example, experimenting with colour shifts), further hyperparameter tuning and noise removal from the train

data set. Additionally, label binning will be implemented. Further analyzing test performances may also provide some insights as to why the models performed the way that they did. In other words, it may be worthwhile to look at the data center origin of the test samples to observe any patterns or biases that negatively or positively affected results. Once this future work is completed, the model will be extended to include prostate biopsies provided by Queen's University and further generalize the model.

ACKNOWLEDGMENTS

We would like to acknowledge the Karolinska Institute and Radboud University for providing the data used in this study as well as hosting this Kaggle challenge. Additionally, we would like to acknowledge the MICCAI 2020 Conference for co-hosting this challenge. Further acknowledgements extend to Kaggle user "iafoss" who provided the competitors of this challenge the foundation to implement the concatenated tile extraction method used in preprocessing the large data.

REFERENCES

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A., "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians* **68**(6), 394–424 (2018).

[2] Committee, C. C. S. A. et al., "Canadian cancer statistics 2019. toronto, on: Canadian cancer society; 2019."

[3] Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., Bostwick, D. G., Evans, A. J., Grignon, D. J., Humphrey, P. A., et al., "Pathologist-level grading of prostate biopsies with artificial intelligence," *arXiv preprint arXiv:1907.01368* (2019).

[4] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G., "Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology* **21**(2), 233–241 (2020).

[5] "Prostate cancer grade assessment (panda) challenge."

[6] Chollet, F. et al., "Keras," (2015).

[7] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X., "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.

[8] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," *Communications of the ACM* **60**(6), 84–90 (2017).