EUROPEAN UROLOGY FOCUS 7 (2021) 347–351

available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Development of a Deep Learning Algorithm for the Histopathologic Diagnosis and Gleason Grading of Prostate Cancer Biopsies: A Pilot Study

Ohad Kott [a,†], *Drew Linsley* [b,†], *Ali Amin* [c,d], *Andreas Karagounis* [b], *Carleen Jeffers* [b], *Dragan Golijanin* [a,d,e], *Thomas Serre* [b,‡], *Boris Gershman* [f,‡,*]

[a] *Minimally Invasive Urology Institute, The Miriam Hospital, Providence, RI, USA;* [b] *Carney Institute for Brain Science, Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI, USA;* [c] *Department of Pathology and Laboratory Medicine, The Miriam Hospital, Providence, RI, USA;* [d] *Warren Alpert Medical School of Brown University, Providence, RI, USA;* [e] *Division of Urology, Rhode Island Hospital and The Miriam Hospital, Providence, RI, USA;* [f] *Division of Urologic Surgery, Beth Israel Deaconess Medical Center, Boston, MA, USA*

Article info

Associate Editor: Derya Tilki

Keywords:
Machine learning
Deep learning
Prostate cancer
Diagnosis
Gleason grade

Abstract

Background: The pathologic diagnosis and Gleason grading of prostate cancer are time-consuming, error-prone, and subject to interobserver variability. Machine learning offers opportunities to improve the diagnosis, risk stratification, and prognostication of prostate cancer.
Objective: To develop a state-of-the-art deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate biopsy specimens.
Design, setting, and participants: A total of 85 prostate core biopsy specimens from 25 patients were digitized at 20× magnification and annotated for Gleason 3, 4, and 5 prostate adenocarcinoma by a urologic pathologist. From these virtual slides, we sampled 14 803 image patches of 256 × 256 pixels, approximately balanced for malignancy.
Outcome measurements and statistical analysis: We trained and tested a deep residual convolutional neural network to classify each patch at two levels: (1) coarse (benign vs malignant) and (2) fine (benign vs Gleason 3 vs 4 vs 5). Model performance was evaluated using fivefold cross-validation. Randomization tests were used for hypothesis testing of model performance versus chance.
Results and limitations: The model demonstrated 91.5% accuracy ($p < 0.001$) at coarse-level classification of image patches as benign versus malignant (0.93 sensitivity, 0.90 specificity, and 0.95 average precision). The model demonstrated 85.4% accuracy ($p < 0.001$) at fine-level classification of image patches as benign versus Gleason 3 versus Gleason 4 versus Gleason 5 (0.83 sensitivity, 0.94 specificity, and 0.83 average precision), with the greatest number of confusions in distinguishing between Gleason 3 and 4, and between Gleason 4 and 5. Limitations include the small sample size and the need for external validation.
Conclusions: In this study, a deep learning-based computer vision algorithm demonstrated excellent performance for the histopathologic diagnosis and Gleason grading of prostate cancer.
Patient summary: We developed a deep learning algorithm that demonstrated excellent performance for the diagnosis and grading of prostate cancer.

† Joint first authors.
‡ Joint senior authors.
* Corresponding author. Division of Urologic Surgery, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA. Tel. +1 617 6673739.
E-mail address: bgershma@bidmc.harvard.edu (B. Gershman).

https://doi.org/10.1016/j.euf.2019.11.003
2405-4569/

1. Introduction

Accurate pathologic diagnosis and Gleason grading of prostate cancer are essential for risk stratification and appropriate management [1,2] but these tasks are time-consuming and subject to substantial interobserver variability [3,4]. Machine learning offers opportunities to improve the diagnosis, risk stratification, and prognostication of prostate cancer through enhanced classification and prediction in a variety of clinical applications [5]. Recent advances in deep learning methods, fueled by increased computing power and the availability of large data sets, have facilitated remarkable progress in the field of computer vision [6,7]. For instance, in the CAMELYON16 challenge, some deep learning algorithms demonstrated better performance than human pathologists at detecting breast cancer metastases in whole-slide images of lymph nodes [8].

Although deep learning algorithms have the potential to improve the diagnosis, Gleason grading, and prognostication of prostate cancer, attempts to do so have been limited [9–14]. Moreover, to the best of our knowledge, only one study used prostate core biopsy specimens for training [9]. Given that initial diagnosis and treatment selection are based on core biopsy pathology [15], a deep learning algorithm to improve diagnosis and Gleason grading specifically in core biopsy specimens would have profound clinical applications. We therefore conducted a pilot study to develop a deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate core biopsy specimens.

2. Patients and methods

2.1. Study cohort

After obtaining institutional review board approval, we identified 25 patients from the Miriam Hospital institutional pathology database who underwent ≥12-core transrectal ultrasound-guided prostate biopsy from January 2011 to November 2012 with a diagnosis of prostate cancer.

2.2. Slide digitization and annotation

A total of 85 prostate core biopsy slides were digitized at 20× magnification using an Aperio ScanScope CS scanner (Leica Biosystems, Nussloch, Germany). Each slide was re-reviewed by a fellowship-trained urologic pathologist, who then annotated the slides using Aperio ImageScope v.12.3 software (Leica Biosystems) for regions of Gleason 3, Gleason 4, and Gleason 5 prostate adenocarcinoma to create pixel-level annotations (Supplementary Fig. 1). Benign patches were sampled from non-cancer-containing regions on the same slides from which cancer-containing patches were sampled to avoid model overfitting on artifactual differences between digitized slides.

2.3. Development and evaluation of the deep learning algorithm

From the 85 virtual slides, we sampled 14 803 image patches of 256 × 256 pixels in size. A patch was considered to contain prostate adenocarcinoma if >60% of the pixels were annotated as such. We then trained an 18-layer-deep residual convolutional neural network (CNN; ResNet) [16] to classify each patch at two levels: (1) coarse classification as benign versus malignant; and (2) fine classification as benign versus Gleason 3 versus Gleason 4 versus Gleason 5. The sample was separated into five training and test sets, with training sets consisting of 80% of the slides (split by unique patients). Models were trained to minimize cross entropy between predicted class probabilities and ground truth labels, and we report performance on predictions from concatenated validation sets.

Model performance was evaluated using fivefold cross-validation over unique patients and is reported as accuracy, sensitivity, specificity, and average precision (weighted area under the precision-recall curve) [17]. Randomization tests were used for hypothesis testing of the model performance against chance [18]. This involved generating a null distribution of the model performance by recalculating model accuracy after shuffling the associations between its predictions and image patch labels. Null distributions consisted of 10 000 such simulations, and p values were calculated as the proportion of simulations that exceeded the true model accuracy.

Models were trained and evaluated using Tensorflow v.1.5 (www.tensorflow.org).

3. Results

Pathologic characteristics for the 85 annotated slides and 14 803 patches of 256 × 256 pixels are presented in Table 1. The CNN was separately trained for patch-based classification as (1) benign versus malignant (coarse classification) and (2) benign versus Gleason 3 versus Gleason 4 versus Gleason 5 (fine classification).

The model demonstrated 91.5% accuracy for coarse classification of image patches as benign versus malignant ($p < 0.001$). This corresponded to sensitivity of 0.93, specificity of 0.90, and average precision of 0.95 (Fig. 1). The AUC was 0.83.

The model demonstrated 85.4% accuracy for fine classification of image patches as benign versus Gleason 3 versus Gleason 4 versus Gleason 5 ($p < 0.001$). This corresponded to sensitivity of 0.83, specificity of 0.94, and average precision of 0.83 (Fig. 2), with the greatest number of confusions in distinguishing between Gleason 3 and 4, and between Gleason 4 and 5.

Table 1 – Baseline characteristics for the study cohort of 25 patients

Feature	*n* (%)
Total patients	25 (100)
Total cores/slides	85 (100)
Slide pathology[a]	
Benign	85 (100)
Any Gleason	85 (100)
Contains Gleason 3[b]	57 (67)
Contains Gleason 4[b]	24 (28)
Contains Gleason 5[b]	25 (29)
Patch pathology	14 803
Benign	6504
Gleason 3	4295
Gleason 4	2784
Gleason 5	1220

[a] Benign patches were sampled from cancer-containing slides outside of annotated cancer regions to avoid model overfitting on artifactual differences between digitized slides.
[b] Gleason score slide totals exceed the total number of slides with prostate adenocarcinoma because a given slide may contain multiple regions with different Gleason scores.

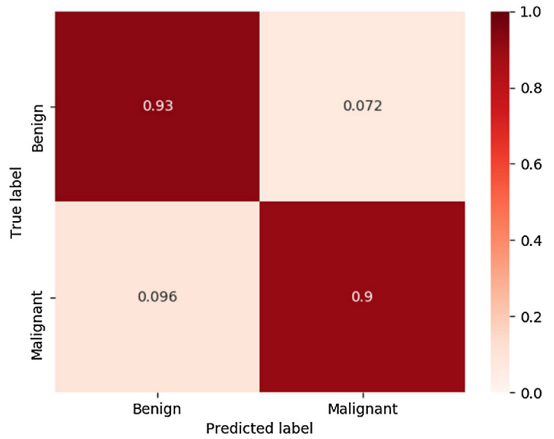

Fig. 1 – Patch-based classification as benign versus malignant (coarse classification).

4. Discussion

The deep learning algorithm developed in this pilot study demonstrated remarkably high accuracy in both coarse (benign vs malignant) and fine (benign vs Gleason 3 vs Gleason 4 vs Gleason 5) classification tasks. Although there was some misclassification of Gleason patterns 3, 4, and 5, as seen in the confusion matrix in Fig. 2, these rates were well within the rate of interobserver variability among human pathologists of 15–30% [3,4]. Moreover, we observed high performance for a relatively small sample size (relative to the size of machine learning data sets).

Few groups have attempted to develop deep learning algorithms for the diagnosis and/or Gleason grading of prostate cancer, almost all using prostatectomy specimens, with modest performance reported [9–14]. Two groups used tissue microarrays derived from radical prostatectomy specimens to develop patch-based deep learning algorithms for prostate cancer diagnosis and Gleason grading [11,13]. Nir et al. [13] reported accuracy of 92% for classification of benign versus malignant and 78% for classification of benign versus low-grade versus high-grade (Gleason 4–5) cancer. Arvaniti et al. [11] reported precision (ie, correct prediction with at least one of two pathologist labels) of 58% for benign patches, 75% for Gleason 3, 86% for Gleason 4, and 58% for Gleason 5. Zhou and colleagues [12] used 380 prostatectomy whole-slide images from The Cancer Genome Atlas (TCGA) to differentiate Gleason 3 + 4 from 4 + 3 with accuracy of 75%. Using one of the largest prostatectomy-based data sets comprising 1226 annotated slides from TCGA, single-institution samples, and an independent medical laboratory, Nagpal and colleagues [14] trained a deep learning algorithm that had a mean accuracy of 70% compared with 61% among 29 general pathologists.

Importantly, to the best of our knowledge, only one other group has trained a deep learning algorithm specifically using prostate core biopsy specimens. Campanella and coauthors [9] used 12 160 whole-slide images from prostate core biopsies to train a semi-supervised deep learning algorithm that had an AUC of 0.98. Their impressive results appear to stand out when compared to the performance of the prostatectomy-based studies discussed above [11–14].

Interestingly, Bartels and colleagues [19–21] reported on the development of a machine vision system for the diagnosis of prostate cancer and identification of cribriform pattern more than 20 years ago. Although in many respects the work was ahead of its time, there are important distinctions compared to contemporary methods in computer vision.

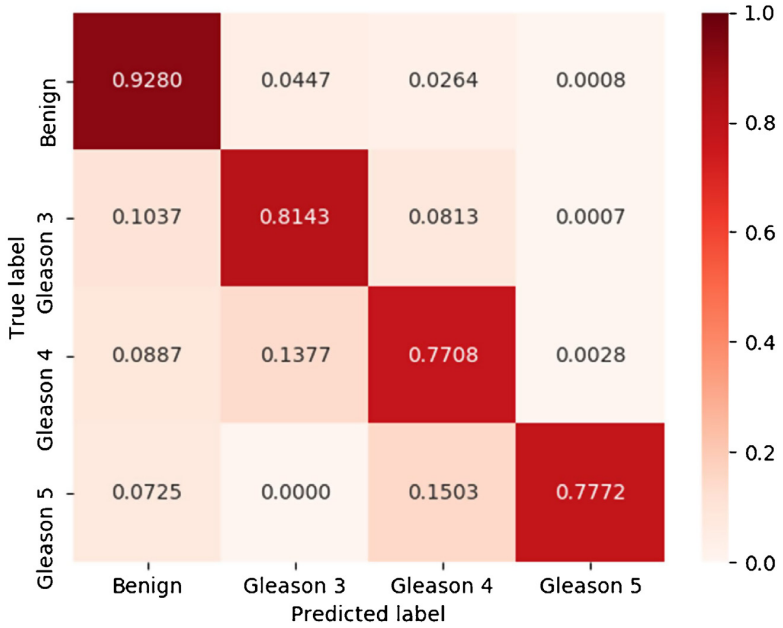

Fig. 2 – Patch-based classification as benign versus Gleason 3 versus Gleason 4 versus Gleason 5 (fine classification).

Specifically, contemporary deep learning approaches are a form of "representation learning" that is entirely data driven rather than relying on individual features engineered by humans, as was common in earlier computer vision approaches [6]. The models are therefore not biased towards selection of particular features and may identify novel features that result in both better performance and application to a diverse array of image classification problems.

The distinction between deep learning algorithms based on prostatectomy specimens and those based on core biopsy is a clinically salient one for many reasons. Specifically, the initial diagnosis, risk stratification, and treatment decisions for men with prostate cancer are based on core biopsy pathology [15]. Accordingly, the performance of algorithms developed for prostatectomy specimens may not directly transfer to core biopsy specimens given the markedly smaller tissue specimen and potential for oblique core sampling to alter histologic architecture.

There are other potential applications of a deep learning algorithm that improves the diagnosis and Gleason grading of prostate core biopsy given its central role in the evaluation and management of prostate cancer. For instance, such an algorithm might expand access to expert pathologic diagnosis not only across the USA but globally to regions where access to high-quality health care may be limited [22]. In settings with established pathologic expertise, such a system could be used to minimize human error as part of quality assurance/improvement efforts. Moreover, deep learning algorithms have the potential not only to recapitulate pathologic diagnosis and contemporary Gleason grading systems but also to discover novel morphological features that are relevant to cancer prediction and prognostication, thereby improving performance.

One important consideration for application of such a deep learning model at other centers relates to image preprocessing. The image patches in this study required preprocessing (ie, to have zero mean unit variance) and use of the model in other centers would require fine-tuning to account for potential differences in such parameters given potential differences in tissue preparations, microscopes used, etc. However, with such adjustments it should be possible, in principle, to analyze any digitized prostate biopsy specimen with the model developed here.

This pilot study has a number of limitations. Foremost, it represents results from a small, single-institution cohort; thus, the algorithm will improve with additional training data and it requires external validation. In addition, the algorithm produces patch-based predictions, although extension to a core-based system would not require substantial technical modifications. Furthermore, the deep learning model was not trained to differentiate specific morphological subtypes of Gleason pattern 4, which may have biological implications. Finally, although each slide was re-reviewed by a urologic pathologist, the study would benefit from multiple experts to generate consensus-based ground truth labels. Despite these limitations, the pilot study provides compelling data supporting the feasibility and utility of a deep learning algorithm for the diagnosis and Gleason grading of prostate core biopsy specimens. Additional studies are currently ongoing to extend these results and examine other clinically relevant outcomes.

5. Conclusions

In this pilot study, a deep learning-based computer vision algorithm demonstrated excellent accuracy for histopathologic diagnosis and Gleason grading of prostate cancer. These results are encouraging for future clinical application of automated histopathologic diagnosis with deep learning.

Author contributions: Boris Gershman had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.*Study concept and design:* Kott, Linsley, Amin, Karagounis, Jeffers, Golijanin, Serre, Gershman.
Acquisition of data: Kott, Amin, Golijanin, Gershman.
Analysis and interpretation of data: Kott, Linsley, Amin, Karagounis, Jeffers, Golijanin, Serre, Gershman.
Drafting of the manuscript: Kott, Linsley, Amin, Karagounis, Jeffers, Golijanin, Serre, Gershman.
Critical revision of the manuscript for important intellectual content: Kott, Linsley, Amin, Karagounis, Jeffers, Golijanin, Serre, Gershman.
Statistical analysis: Linsley, Karagounis, Jeffers, Serre, Gershman.
Obtaining funding: Kott, Linsley, Golijanin, Serre, Gershman.
Administrative, technical, or material support: None.
Supervision: Gershman, Serre.
Other: None.

Financial disclosures: Boris Gershman certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: Thomas Serre serves on the scientific advisory board for Vium Inc. The remaining authors have nothing to disclose.

Funding/Support and role of the sponsor*:** This study was supported by NIGMS/Advance-CTR through IDeA-CTR grant U54GM115677. Additional support was provided by the Carney Institute for Brain Sciences, the Center for Vision Research, and the Center for Computation and Visualization. The sponsors played no direct role in the study.Acknowledgments:*** We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.euf.2019.11.003.

References

[1] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40:244–52.
[2] Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. Eur Urol 2016;69:428–35.

[3] Allsbrook Jr WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. Hum Pathol 2001;32:81–8.
[4] Sauter G, Steurer S, Clauditz TS, et al. Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. Eur Urol 2016;69:592–8.
[5] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58.
[6] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[7] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. 2017 https://arxiv.org/abs/1707.02968
[8] Ehteshami Bejnordi B, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.
[9] Campanella G, Werneck Krauss Silva V, Fuchs TJ. Terabyte-scale deep multiple instance learning for classification and localization in pathology. 2018 https://arxiv.org/abs/1805.06983
[10] Nir G, Karimi D, Goldenberg SL, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. JAMA Netw Open 2019;2:e190442.
[11] Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep 2018;8:12054.
[12] Zhou N, Fedorov A, Fennessy F, Kikinis R, Gao Y. Large scale digital prostate pathology image analysis combining feature extraction and deep neural network. 2017 https://arxiv.org/abs/1705.02678
[13] Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. Med Image Anal 2018;50:167–80.
[14] Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Dig Med 2019;2:48.
[15] Mohler JL, Armstrong AJ, Bahnson RR, et al. Prostate Cancer, Version 1.2016. J National Compr Cancer Network : JNCCN 2016;14(1):19–30.
[16] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. 2016 https://arxiv.org/abs/1603.05027
[17] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) challenge. Int J Comput Vis 2009;88:303–38.
[18] Edgington ES. Randomization tests. J Psychol 1964;57:445–9.
[19] Bartels PH, Bartels HG, Montironi R, Hamilton PW, Thompson D. Machine vision in the detection of prostate lesions in histologic sections. Anal Quant Cytol Histol 1998;20:358–64.
[20] Bartels PH, Thompson D, Bartels HG, Montironi R, Scarpelli M, Hamilton PW. Machine vision-based histometry of premalignant and malignant prostatic lesions. Pathol Res Pract 1995;191:935–44.
[21] Thompson D, Bartels PH, Bartels HG, Montironi R. Image segmentation of cribriform gland tissue. Anal Quant Cytol Histol 1995;17:314–22.
[22] Chen P-HC, Gadepalli K, MacDonald R, et al. Microscope 2.0: an augmented reality microscope with real-time artificial intelligence integration. 2018 https://arxiv.org/abs/1812.00825