A Hybrid Human–Machine Learning Approach for Screening Prostate Biopsies Can Improve Clinical Efficiency Without Compromising Diagnostic Accuracy

David Dov, PhD; Serge Assaad, BS; Ameer Syedibrahim; Jonathan Bell, MD; Jiaoti Huang, MD, PhD; John Madden, MD, PhD; Rex Bentley, MD; Shannon McCall, MD; Ricardo Henao, PhD; Lawrence Carin, PhD; Wen-Chi Foo, MD

- **Context.**—Prostate cancer is a common malignancy, and accurate diagnosis typically requires histologic review of multiple prostate core biopsies per patient. As pathology volumes and complexity increase, new tools to improve the efficiency of everyday practice are keenly needed. Deep learning has shown promise in pathology diagnostics, but most studies isolate the efforts of pathologists from the application of deep learning algorithms. Very few hybrid pathologist–deep learning approaches have been explored, and these typically require complete review of histologic slides by both the pathologist and the deep learning system.

- **Objective.**—To develop a novel and efficient hybrid human–machine learning approach to screen prostate biopsies.

- **Design.**—We developed an algorithm to determine the 20 regions of interest with the highest probability of malignancy for each prostate biopsy; presenting these regions to a pathologist for manual screening limited the initial review by a pathologist to approximately 2% of the tissue area of each sample. We evaluated this approach by using 100 biopsies (29 malignant, 60 benign, 11 other) that were reviewed by 4 pathologists (3 urologic pathologists, 1 general pathologist) using a custom-designed graphical user interface.

- **Results.**—Malignant biopsies were correctly identified as needing comprehensive review with high sensitivity (mean, 99.2% among all pathologists); conversely, most benign prostate biopsies (mean, 72.1%) were correctly identified as needing no further review.

- **Conclusions.**—This novel hybrid system has the potential to efficiently triage out most benign prostate core biopsies, conserving time for the pathologist to dedicate to detailed evaluation of malignant biopsies.

(Arch Pathol Lab Med. doi: 10.5858/arpa.2020-0850-OA)

Prostate cancer is one of the most common malignancies in the developed world and is projected to account for more than 190,000 new cancer diagnoses in the United States this year. The gold standard for prostate cancer diagnosis involves analysis of a series of prostate needle core biopsies, which represent the majority of samples in most genitourinary pathology practices. Rendering a diagnosis for prostatic adenocarcinoma—the most common prostate malignancy, accounting for 95% of prostate cancers—is a multistep, time-consuming process. If adenocarcinoma is found to be present in a prostate biopsy, subsequent steps in the diagnosis consist of determining the type of adenocarcinoma (acinar versus ductal), followed by Gleason grade, prognostic Grade Group, and quantity. In intermediate-grade tumors, the percentage of Gleason pattern 4 disease is also reported. Furthermore, recent guidelines recommend the additional step of reporting the presence or absence of cribriform architecture. Clearly, as pathology volumes and the complexity of cases increase, the creation of tools to improve the efficiency of everyday practice would be invaluable. Deep learning is one such tool that has shown tremendous promise.

Several previous studies have applied deep learning approaches to prostate needle core biopsies. Most have involved building algorithms with preanalytic requirements for intensive manual annotation of slides by pathologists. In contrast, Campanella et al pioneered a weakly supervised approach that obviates the need for time-consuming manual annotation of slides. Instead, using only specimen-level diagnoses as whole-slide labels, a highly accurate algorithm was developed that recognized the presence of cancer on a whole-slide level. The authors proposed a diagnostic model in which all prostate biopsies are initially screened by their algorithm and those deemed benign need never be reviewed by a pathologist at all. Strikingly, when they tested this model, they found that their algorithm performed well enough to not miss any cancer at a patient level (ie, at least 1 slide was called malignant for every patient who truly had prostate cancer). The algorithm of Campanella and colleagues shows...
substantial promise, yet it also raises philosophical questions regarding the practice of pathology and the limits of deep learning.

Many pathologists feel apprehensive about the notion of assigning benign diagnoses to biopsy specimens without any human review. Indeed, because deep learning is essentially empirical, the accuracy of an algorithm is highly dependent on the number of examples on which it is trained. While 95% of cancers arising in the prostate are acinar adenocarcinomas with similar morphologic features, many variants of prostate adenocarcinoma exist, as do other types of prostate cancer such as small cell carcinoma and basal cell carcinoma. Other neoplasms that are not necessarily of prostatic epithelial origin, such as lymphoma, urothelial carcinoma, and stromal tumors, can also be encountered. These rare cases are unlikely to be present in sufficient frequency in deep learning training sets to be accurately identified; thus, using deep learning to definitively triage prostate biopsies and entirely omit a subset of specimens from pathologist review would seem potentially problematic. In addition, there are inherent limitations to the practical application of machine learning in a medical setting: changes in practice guidelines, processing protocols, upgrading of instruments, among others, can all impact the accuracy of an algorithm. Therefore, responsible use of deep learning requires constant monitoring at a human level.

To address these concerns, we aimed to create a novel hybrid diagnostic process in which a deep learning algorithm is used in conjunction with pathologist expertise for all cases. Drawing on the example of a common cytopathology workflow for triaging Papanicolaou (Pap) smears, we developed a weakly supervised deep learning algorithm to identify a predetermined number of regions of interest (ROIs), highlighting the most suspicious areas from each biopsy for the pathologist to personally review. If any ROI was deemed malignant, or even equivocal for malignancy, that biopsy was categorized as needing comprehensive manual review; if all ROIs were deemed benign, that biopsy was categorized as needing no further review. With this protocol, we determined that this model results in very high sensitivity and interobserver agreement in identifying specimens that require comprehensive manual examination. When applied to an entirely digitized pathology system, our model may allow pathologists to sign out negative prostate needle core biopsies after examining only a fraction of the total area of the tissue, improving efficiency and allowing most time and effort to be dedicated to clinically significant atypical or malignant slides.

**MATERIALS AND METHODS**

This study was deemed exempt by our hospital’s institutional review board. From our hospital’s electronic records we retrieved all prostate needle core biopsy reports verified from January 18, 2018, through March 19, 2019. This initially represented a total of 820 cases verified by 6 different attending pathologists, 3 of whom specialize in genitourinary pathology. These 3 urologic pathologists verified between 112 and 301 cases each (mean, 216 cases), while the pathologists without an expertise in genitourinary pathology verified between 23 and 76 cases each (mean, 50 cases). Cases were retrieved from the pathology archives in chronologic order with an effort to include cases verified by all 6 attending pathologists. All clinical cases were composed of 3 slides stained with hematoxylin–eosin per biopsy. No biopsies were excluded by specific diagnosis, treatment history, tissue artifact (ie, folds, fragmentation, tissue tears), or preparation artifact (differential staining, slide bubbles).

After the cases were retrieved from the pathology archives, they were re-reviewed microscopically by one of the pathologists with genitourinary expertise. If on re-review the diagnosis matched the original diagnosis, the original pathologists’ whole biopsy-level diagnoses were used as whole-slide–level labels. One hematoxylin–eosin–stained representative slide from each biopsy was selected for scanning. Slides were scanned with a whole slide scanner (Aperio AT2, Leica Biosystems) at ×40 magnification, and de-identified. For time efficiency, slide scanning was arbitrarily halted after 2123 biopsies were reviewed. Re-review produced diagnostic disagreements on 11 biopsies, which were excluded from the study for simplicity. These disagreements were all between diagnoses of benign and atypical small acinar proliferation (ASAP); no clinically significant disagreements were identified. Because machine learning algorithms are considered robust to a small amount of label noise, these 11 biopsies out of the 2123 total would unlikely have meaningfully affected the algorithm regardless of inclusion or exclusion. Additionally, for simplicity, 61 scanned needle core biopsies where the biopsy tissue spanned over more than 1 slide were also excluded. After these exclusions, the dataset comprised 2051 needle core biopsies from 172 patient cases. The number of cases verified by each attending ranged from 23 to 39. Most cases were composed of 12 needle core biopsies, but the number of biopsies per case ranged from 4 to 26, with a mean of 11.9. These needle core biopsies were then divided at the patient level into a training set (1257 slides), validation set (391 slides), and test set (403 slides). See Table 1 for details regarding clinical distribution of diagnoses.

To build a program to automatically identify ROIs for evaluation by a pathologist, we first designed a binary classifier to predict malignancy at the slide level. A binary classifier (rather than multiclass, which could include multiple nonmalignant diagnoses) was chosen so that binary predictions could serve as a proxy to the importance of the ROIs, allowing the ROIs with the highest prediction values for malignancy to be selected for pathologist review. Based on their diagnoses, biopsies were dichotomized as either “malignant” or “nonmalignant” (the latter encompassing benign biopsies as well as ASAP and high-grade prostatic intraepithelial neoplasia [HGPIN]). The decision to group ASAP and HGPIN together with benign specimens (rather than grouping them with malignant specimens and thereby dichotomizing biopsies as “benign” or “nonbenign”) was made in order to build an algorithm designed to minimize false negatives (ie, not miss malignancies). Defining the biopsies as “malignant” versus “nonmalignant” allowed for enrichment of definitive features of malignancy in the training set without dilution by atypical but not outright malignant features as seen in ASAP or HGPIN.

To build the binary classifier, we used multiple instance learning (MIL), a learning approach where the slide is divided and processed in multiple patches. Each patch was fed into a neural network, and in turn, the patch-level prediction scores were aggregated into a global, slide-level prediction. Specifically, we used the noisy-or approach, in which the global prediction is simply given by the maximal patch-level prediction. The motivations for this approach were 3-fold: (1) the patch-based processing allows us to overcome memory limitations posed by the large size of the slides; (2) MIL allows using only slide-level labels for training without the need to annotate individual patches; and (3) we hypothesized that the MIL algorithm, previously shown to be useful for binary classification of prostate core biopsies, would provide high prediction values to malignant patches even though it was trained by using the slide-level labels only.

To train the classifier, the slides were processed in nonoverlapping patches of 600 × 600 pixels in size. Before feeding the patches to the network, they were downsampled by a factor of 2. Patches of this size and resolution were previously used by Campanella et al for binary classification. We also filtered out irrelevant patches comprising white area, using a simple, yet effective approach based on thresholding. Specifically, we took the mean of all the pixel values in each patch and compared it to an upper and lower threshold. If the mean value was between 135 and 220 (on a scale
Table 1. Distribution of Diagnoses for Biopsies in the Training, Validation, and Test Sets

| Diagnosis                        | No. of Biopsies |
|----------------------------------|----------------|
|                                  | Training Set (N = 1257) | Validation Set (N = 391) | Test Set (N = 403) |
| Benign                           | 878            | 226            | 255            |
| Malignant                        |                |                |                |
| Grade Group 1 prostatic adenocarcinoma | 103            | 42             | 35             |
| Grade Group 2 prostatic adenocarcinoma | 61             | 27             | 28             |
| Grade Group 3 prostatic adenocarcinoma | 46             | 21             | 19             |
| Grade Group 4 prostatic adenocarcinoma | 20             | 7              | 7              |
| Grade Group 5 prostatic adenocarcinoma | 25             | 23             | 17             |
| Other                            |                |                |                |
| ASAP                             | 36             | 14             | 22             |
| HGPIN                            | 88             | 31             | 20             |

Abbreviations: ASAP, atypical small acinar proliferation; HGPIN, high-grade prostatic intraepithelial neoplasia.

between 0 and 255), then we selected it as part of the tissue. Then, we used a convolutional neural network based on the VGG-11 architecture13 with convolutional filters pretrained on the ImageNet dataset.14 During training, we kept the convolutional filters fixed and only tuned the fully connected layers, significantly reducing the training time. The network was trained for 10 epochs, using the binary cross entropy loss with the slide-level labels. We used the Adam optimizer15 with the learning rate of 1e-3 and weight decay parameters were saved upon improvement.

The binary MIL-based classifier was able to accurately determine whether a prostate biopsy was benign or malignant, providing an area under the curve of 0.961 over the test set (Figure 1). This binary classifier model was then used to create a program that ranked the patches comprising each prostate needle core by the probability of malignancy. An arbitrary decision was made to show the 20 most suspicious patches (defined as the ROIs) to the pathologist; these 20 patches were those for which the neural network provided the highest malignancy prediction values. The ROIs were presented (in full resolution) for review via a custom software that we designed. A screenshot is demonstrated in Figure 2.

From the test set, 100 biopsies were randomly selected, which included 60 benign biopsies, 29 malignant biopsies, and 11 biopsies diagnosed as “other” (either ASAP or HGPIN; see Table 2 for detailed distribution of clinical diagnoses). A pathologist with genitourinary expertise reviewed the 20 most suspicious ROIs from each biopsy in a blinded manner. For each biopsy, individual ROIs were first characterized by the pathologist as benign, equivocal for malignancy, or malignant. The presence of the “equivocal” category allows the pathologist to flag a case as needing further review if findings are atypical, even if morphologic features seen in a specific ROI fall short of a definitive diagnosis of malignancy. This category was created to allow increased sensitivity in identifying potentially malignant biopsies.

After evaluation of all 20 ROIs from each biopsy, an overall biopsy-level assignment was made, based on the worst overall ROI characterization. If all 20 ROIs were deemed benign, the biopsy was given an overall slide-level designation as benign. If any single ROI was deemed malignant, the entire biopsy was given a slide-level designation as malignant. If a biopsy had at least 1 ROI deemed equivocal, but none deemed malignant, then the entire biopsy was given a slide-level designation as equivocal. All biopsies with an overall slide-level designation of “equivocal” or “malignant” were categorized as biopsies needing further review (Figure 3). This pathologist’s individual ROI characterization and overall slide-level designations were analyzed and compared with the clinical overall slide-level diagnoses.

In the subsequent reproducibility study, ROI evaluation was performed by 3 additional pathologists, 2 of whom also concentrate on genitourinary pathology. The third pathologist is a senior general pathologist without a specific concentration in genitourinary pathology. Overall slide-level designations from these additional pathologists were recorded, based on worst overall ROI characterization, compared with the clinical overall slide-level diagnoses and also analyzed as above.

RESULTS

Each set of 20 ROIs represented, on average, 2% of the total surface area of tissue per slide. In the initial phase of the study, a single pathologist with genitourinary expertise screened a total of 2000 ROIs taken from 100 whole slide images of needle core biopsies, and 29 of 29 biopsies (100%) clinically diagnosed as malignant were identified as needing further review through ROI analysis (ie, at least 1 ROI was deemed “equivocal” or “malignant”). This represents a sensitivity of 100% for detecting biopsies with cancer through ROI analysis. Conversely, of the 71 nonmalignant biopsies, 60 were clinically diagnosed as outright benign; 50 (83.3%) of these benign biopsies were scored as needing no further review on ROI analysis (Tables 3 and 4, pathologist 1). Of the 100 biopsies, 20 were characterized by the pathologist through ROI analysis as equivocal for malignancy. Of these “equivocal” biopsies, 3 had been clinically diagnosed as having prostate cancer, 10 as benign, and 7 as “other” (5 contained HGPIN and 2 diagnosed as ASAP).
The determinative ROI (ie, the first ROI deemed “equivocal” or “malignant” in those biopsies receiving an overall slide-level designation other than benign) was identified after review of a mean of 3.3 ROIs. Therefore, an average cumulative review of less than 0.5% of the total area of the tissue per slide was required to determine that a biopsy would require detailed review by this pathologist.

In the reproducibility phase of the study, 3 additional pathologists reviewed the same sets of ROIs from the 100 biopsies (Tables 3 and 4); two of these pathologists also have genitourinary expertise, while the third is a general pathologist. Two of the 3 identified 29 of 29 (100%) malignant biopsies as requiring additional review on ROI analysis, and 1 pathologist identified 28 of 29 (96.6%) malignant biopsies as needing further review. Taken together among all 4 pathologists, a mean sensitivity of 99.2% was achieved for the malignant needle core biopsies (Table 5). Of the 60 biopsies clinically diagnosed as benign, 26 to 55 (43.3%–91.7%) were scored as benign on ROI analysis by the 3 additional pathologists. The general pathologist scored 26 (43.3%) of the clinically diagnosed benign biopsies as benign, whereas the genitourinary pathologists designated 42 to 55 (70%–91.7%) of the benign biopsies as benign. Altogether, an average of 72.1% of benign biopsies were designated as not requiring further review (Table 5).

To assess the level of efficiency of screening, the determinative ROI for biopsies deemed equivocal or malignant was determined for all pathologists. On average, only 4.4 of 20 ROIs were reviewed before a pathologist designated a biopsy as needing further review. In biopsies clinically diagnosed as malignant, more than 80% of cases were identified as needing further review after evaluation of only a single ROI. In all cases the ultimate designation of a biopsy as “equivocal” or “malignant” was made before the 20th ROI (Figure 4).

Breaking down the specimen ROI determinations for biopsies ultimately characterized as malignant, there was a mean of 10.4 ROIs characterized as “malignant,” 5.1 ROIs characterized as “equivocal,” and 4.6 ROIs characterized as “benign.” For the biopsies ultimately deemed equivocal on ROI analysis, there was a mean of 2.7 ROIs characterized as “equivocal” and 17.3 ROIs characterized as “benign.” By definition, all biopsies ultimately deemed “benign” on ROI analysis had all 20 ROIs characterized as benign.

**DISCUSSION**

Deep learning has shown promise in improving diagnostic accuracy and/or efficiency in a number of medical fields, including radiology, ophthalmology, and pathology. Multiple previous studies have specifically examined the applicability of deep learning to the diagnosis of prostate cancer based on histomorphology. Many have centered on prostatectomy specimens, either through examination of whole prostatectomy slides or tissue microarrays produced from prostatectomy tissue. While studies on prostatectomy specimens have demonstrated accuracy in recognizing prostate cancer and even in assessing Gleason grade on a whole-slide level, a pathologist’s review of prostatectomy cases is substantially more complex than this, requiring synthesis of information over many slides as well as assessment of additional features for tumor staging. Deep learning might be more productively applied to prostate needle core biopsies, which account for a large proportion of genitourinary pathology cases. In current clinical practice, when no cancer is identified within a particular biopsy, it can simply be deemed benign with no additional informa-
Figure 3. Diagram of project workflow. After the prostate biopsies were obtained, they were scanned at ×20 magnification to give rise to whole slide images (WSIs). The machine learning algorithm was applied to each WSI, and the 20 most suspicious areas (regions of interest [ROIs]) were selected. These ROIs were showcased in a custom-designed graphical interface for evaluation by the pathologists, who—from their reviews of the ROIs—determined whether a prostate biopsy needed further review.

As our system was used to develop a system of efficiently excluding benign prostate core biopsies through a hybrid approach using deep learning in combination with a pathologist’s expertise. Hybrid models combining human expertise with computer algorithms are not a new concept: commercially available systems are currently in use in the area of Pap smear cytology. Currently, Pap smear specimens are predominantly processed by liquid-based preparations, such as ThinPrep (Hologic, Inc) and SurePath (Becton, Dickinson and Company), and imaging systems linked to these preparations have been developed to aid in their screening. One example is the ThinPrep Imaging System (Hologic), which scans an entire slide and highlights the 22 fields of view with the most suspicious features, as selected via a proprietary algorithm. A pathologist or cytotechnologist assesses these preselected fields of view; if all of these fields are normal, the case can be signed out directly as negative without review of the remainder of the slide. Only if an abnormality is noted within any of the 22 fields is the entire slide manually reviewed. These types of imaging systems have been shown not only to increase efficiency in workflow, but also to improve diagnostic accuracy in large-scale studies.

Using the ThinPrep Pap workflow as inspiration, we built and used a deep learning system to identify 20 ROIs—the 20 most suspicious areas on a slide—for each prostate biopsy and compile them for review by the pathologists in the study. We found that by applying this hybrid system in a digital workflow, whole slide images of malignant prostate needle core biopsies could be triaged as needing further review with almost 100% accuracy. Three of 4 pathologists identified 29 of 29 malignant biopsies as requiring further evaluation, and the fourth identified 28 of the 29 malignant biopsies accordingly. The pathologist who missed 1 malignant biopsy was the only participant who performed the review via a remote desktop telecommunication application owing to COVID-19 (coronavirus disease 2019) pandemic constraints. The remaining pathologists performed their reviews directly in an office setting.

Importantly, an average of 72.1% of whole slide images of benign biopsies were correctly identified as having no areas of concern after review of the 20 most suspicious ROIs per specimen. One of the pathologists (a general pathologist) identified benign biopsies as needing no further review in 43.3% of cases; the other 3 pathologists (with specialization in urologic pathology) correctly identified benign biopsies in 70% to 91.7% of cases (mean, 81.7%). Therefore, while our system might allow a general pathologist to directly verify a sizeable fraction of benign biopsies after reviewing only 2% of the tissue area, this rate may be even higher for urologic pathologists who review a greater proportion of prostate biopsies in their everyday practice. Given that most prostate biopsies submitted are in fact benign, the potential time savings of triaging benign biopsies in this manner could greatly improve the efficiency of clinical workflow for both general and genitourinary pathologists. While our current study focused on reducing the total area of each biopsy that needed to be reviewed by a pathologist, future studies comparing timed assessments of whole slide image review to ROI-based triage may provide enlightening data with regard to workflow efficiency.

The decision to use 20 ROIs for review instead of a different number was arbitrarily made, but is similar to the 22 used by the ThinPrep Imaging System for Pap smears. Importantly, when the specific ROI designations were examined in malignant biopsies, we found that in most cases (more than 80%), malignant biopsies were designated as requiring further evaluation after the pathologist had

| Clinical Diagnosis (WSI-Based) | Pathologist 1 | Pathologist 2 | Pathologist 3 | Pathologist 4 |
|-------------------------------|---------------|---------------|---------------|---------------|
|                                | Benign | Equivocal | Malignant | Benign | Equivocal | Malignant | Benign | Equivocal | Malignant | Benign | Equivocal | Malignant |
| Benign                         | 50    | 10        | 0         | 55    | 3         | 2         | 26     | 32        | 2         | 42     | 16        | 2         |
| Other                          | 2     | 7         | 2         | 6     | 2         | 3         | 1      | 7         | 3         | 5      | 3         | 3         |
| Malignant                      | 0     | 3         | 26        | 1     | 2         | 26        | 0      | 2         | 27        | 0      | 1         | 28        |

Abbreviation: WSI, whole slide image.
reviewed a single ROI. The mean number of ROIs needing to be reviewed in biopsies ultimately deemed nonbenign was 4.4, and in no cases did all 20 ROIs need to be assessed before designating a biopsy as equivocal or malignant. This implies that 20 ROIs are sufficient to screen prostate biopsies for efficient triage of benign and malignant cases, although further studies in which the number of ROIs is varied could potentially optimize this number further.

Most previous studies applying deep learning to prostate cancer diagnostics have proposed models with deep learning systems that perform entirely separately from the pathologist, while their development requires extensive manual annotation of slides by pathologists. This may be problematic owing to several reasons. First, the practice of pathology is not static: for instance, while Gleason grading has been considered standard of care for decades, new guidelines regarding its interpretation and application are published every few years. Second, laboratory practice is also not static: periodic upgrades of staining instruments and chemical formulations are a fact of life. Both of these types of changes can and will impact the accuracy of deep learning algorithms longitudinally. Whereas a pathologist can assimilate and apply new recommendations based on evidence-based medicine and readily acclimate to new staining protocols, deep learning systems cannot and will therefore require periodic re-tuning. Extensive preanalytic manual annotation of slides to develop (and then again to re-tune) deep learning systems is an onerous process that may be unnecessary. Two relatively recent studies bear this out.

With increasing recognition of these potential limitations of applying deep learning to independent review of pathology slides, a synergistic approach combining pathologist expertise with deep learning algorithms is now gaining interest. Two recent studies used weakly supervised approaches to build deep learning algorithms to help pathologists identify prostate cancer and even improve Gleason grading. Raciti and colleagues used Paige Prostate Alpha (PPA; a prostate cancer detection system based on the weakly supervised deep learning algorithm first created by Campanella et al) to review all prostate needle core biopsies, characterize each as benign or malignant, and if malignant, mark the most suspicious area. They found that when pathologists reviewed the same set of cores with and without feedback from PPA, the use of PPA increased pathologists’ sensitivities for detecting prostate cancer. However, the clinical workflow for this model does require that a pathologist completely review every slide that PPA reviews first. Interestingly, there was actually a drop in the number of reclassified cores that the pathologists had initially classified correctly without the aid of PPA when those cores were re-reviewed with the aid of PPA; overall sensitivity improved, however, because there was greater improvement in the percentage of reclassified cores the pathologists had initially classified incorrectly, when re-reviewed with PPA.

Bulten and colleagues focused on using their deep learning system to improve interobserver reproducibility in Gleason grading and found that pathologists assisted by their artificial intelligence system exhibited significant improvement in Gleason grading as measured through agreement with an expert reference standard. Not surprisingly, incorporating input from the deep learning system also resulted in less disagreement between pathologists. This model, like that of Raciti et al, requires that a pathologist still review every slide the artificial intelligence system reviews first. A diagnostic model that might improve clinical efficiency and also interobserver variability in Gleason grading (a known area of concern in prostate pathology) would be to first run all digitized prostate slides through ROI analysis as performed in our study and then to subject biopsies designated as requiring further review through a Gleason grading algorithm such as that of Bulten et al.

Our study has several limitations. The current ROI-pathologist user interface highlighted 20 individual ROIs of 600 × 600 pixels for review outside of the context of the whole slide image (Figure 5, A through J). This size was selected to comply with the binary classifier used in Campanella et al and was not specifically optimized for pathologists’ review. Increasing the size of each ROI and/or highlighting the ROI while keeping it within the context of the original whole slide image may further improve the review process by providing additional architectural information to the pathologist. In addition, the 100 specimens reviewed by the pathologists represented only 9 independent patients, as we attempted to emulate true clinical workflow by including all biopsies from each patient for review. It should be noted, though, that while the number of patients is low, the model performs evaluations on a per

### Table 4. Biopsies Assessed as Needing Further Review by Pathologist as Compared With Overall Clinical Diagnosis

| Clinical Diagnosis (WSI-Based) | Pathologist 1 | Pathologist 2 | Pathologist 3 | Pathologist 4 | Mean |
|-------------------------------|--------------|--------------|--------------|--------------|------|
|                               | No Further Review | Full Review | No Further Review | Full Review | No Further Review | Full Review | No Further Review | Full Review | No Further Review | Full Review | No Further Review | Full Review |
| Nonmalignant *                | 52           | 19           | 61           | 10           | 27           | 44           | 47           | 24           |
| Malignant                     | 0            | 29           | 1            | 28           | 0            | 29           | 0            | 29           |

**Abbreviations:** ROI, region of interest; WSI, whole slide image.

*Nonmalignant: benign, high-grade prostatic intraepithelial neoplasia or atypical small acinar proliferation.

### Table 5. Clinical Utility of Region of Interest Analysis of Prostate Biopsies

|                  | Pathologist 1 | Pathologist 2 | Pathologist 3 | Pathologist 4 | Mean |
|------------------|--------------|--------------|--------------|--------------|------|
| Sensitivity of identifying malignant biopsies needing further review, % | 100          | 96.6         | 100          | 100          | 99.2 |
| Percentage of benign biopsies identified as needing no further review, % | 83.3         | 91.7         | 43.3         | 70.0         | 72.1 |
biopsy (and not per patient) level. Future multi-institutional studies using greater numbers of biopsies/independent patients and multiple laboratories processing tissue will test the generalizability of our results.

Using this novel hybrid human–machine learning approach as a starting point, there are numerous additional directions to be explored in future studies. For instance, by adding examples of rare neoplasms that can be found in the prostate other than prostatic adenocarcinoma, the ability of the algorithm to identify features of nonadenocarcinoma neoplasms when selecting ROIs could be tested. This would be important to ensure that uncommon neoplasms would not be missed during triage. If features of nonadenocarcinoma neoplasms were not reliably selected for ROIs, then training an algorithm based on a "benign" versus "non-benign" dichotomy (see Materials and Methods) might prove necessary. This might also improve detection of ASAP and HGPIN, which may be of clinical importance for patients who do not otherwise have evidence of malignancy in their biopsies. Additionally, while the current study used only 1 of 3 tissue levels of each specimen for analysis, in clinical practice, pathologists review multiple levels. Therefore, it would be interesting to determine whether identifying the 20 most suspicious ROIs obtained across multiple levels might further improve the efficiency of distinguishing biopsies requiring comprehensive review from benign biopsies appropriately signed out at triage.

Finally, using the ROI strategy to train a machine learning algorithm to recognize different Gleason patterns, and potentially sort malignant biopsies by these patterns, is another potential avenue of exploration.

In summary, we believe that our novel hybrid deep learning approach combined with pathologist expertise has the potential to provide an efficient and accurate method of triaging prostate needle-core biopsies to improve pathologist workflow.

The authors acknowledge the Duke BioRepository and Precision Pathology Center for whole slide scanning services and the Duke Plus Data Science Initiative for initiating collaboration between the Department of Pathology and Department of Electrical and Computer Engineering.

References

1. Cancer Stat Facts: Prostate Cancer. National Cancer Institute Surveillance, Epidemiology, and End Results Program. 2020. https://seer.cancer.gov/statfacts/html/prost.html. Accessed October 2, 2020.
2. Litjens G, Sanchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016; 6:26286.
3. Lucas M, Jansen I, Savci-Heijink CD, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. Virchows Arch. 2019;475(1):77–83.
4. Kott O, Linsley D, Amin A, et al. Development of a deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate cancer biopsies: a pilot study. Eur Urol Focus. 2021;7(2):S47-S51.
5. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–1309.
6. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21(2):233–241.
7. Strom P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21(2):222–232.
8. Bulten W, Balkenhol M, Belinga JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod Pathol. 2021;34(3):660–671.

9. Raciti P, Sue J, Ceballos R, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Mod Pathol. 2020;33(10):2058–2066.

10. Steiner DF, Nagpal K, Sayres R, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. JAMA Netw Open. 2020;3(11):e2023267.

11. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019;54:280–296.

12. Campapella G, Silva VWK, Fuchs TJ. Terabyte-scale deep multiple instance learning for classification and localization in pathology. Preprint. Posted online May 17, 2018. arXiv:180506983 [cs].

13. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR. 2014;abs/1409.1556.

14. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. Meeting: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE; 2009:248–255.

15. Kingma DP, Ba J. ADAM: a method for stochastic optimization. Preprint. Posted online December 22, 2014. arXiv:1409.1556 [cs].

16. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

17. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med. 2019;2:48.

18. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep. 2016;6(1):12054.

19. 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(3):e190442.

20. Karimi D, Nir G, Fazli L, Black PC, Goldenberg L, Salcudean SE. Deep learning-based Gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE J Biomed Health Inform. 2020;24(5):1413–1426.

21. Li J, Sama KV, Chung Ho K, Gertzch A, Knudsen BS, Arnold CW. A multiscale U-Net for semantic segmentation of histological images from radical prostatectomies. AMIA Annu Symp Proc. 2018;2017:1140–1148.

22. Abraham NE, Mendhiratta N, Taneja SS. Patterns of repeat prostate biopsy in contemporary clinical practice. J Urol. 2015;193(4):1178–1184.

23. Schledermann D, Hyldebrandt T, Ejesbo D, Hoelund B. Automated screening versus manual screening: a comparison of the ThinPrep imaging system and manual screening in a time study. Diagn Cytopathol. 2007;35(6):348–352.

24. Wilbur DC, Black-Schafer WS, Luff RD, et al. The Becton Dickinson FocalPoint GS Imaging System: clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. Am J Clin Pathol. 2009;132(5):767–775.

25. Epstein JJ, Egevad L, Amin MB, et al. 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(2):244–252.

26. Epstein JJ, Amin MB, Fine SW, et al. The 2019 Genitourinary Pathology Society (GUPS) white paper on contemporary grading of prostate cancer. Arch Pathol Lab Med. 2021;145(4):461–493.

27. van Leenders G, van der Kwast TH, Grignon DJ, et al. The 2019 International Society of Urological Pathology (ISUP) Consensus Conference on Grading of Prostatic Carcinoma. Am J Surg Pathol. 2020;44(8):e87–e99.

28. Sadimini ET, Khani F, Diolombi M, Meliti A, Epstein JI. Interobserver reproducibility of percent Gleason pattern 4 in prostatic adenocarcinoma on prostate biopsies. Am J Surg Pathol. 2016;40(12):1686–1692.

29. Nakay T, Tanaka N, Miyake M, et al. Atypical small acinar proliferation and two or more cores of high-grade intraepithelial neoplasia on a previous prostate biopsy are significant predictors of cancer during a transperineal template-guided saturation biopsy aimed at sampling one core for each 1 mL of prostate volume. Res Rep Urol. 2017;9:187–193.