Don't Fear the Artificial Intelligence

A Systematic Review of Machine Learning for Prostate Cancer Detection in Pathology

Aaryn Frewing; Alexander B. Gibson; Richard Robertson; Paul M. Urie, MD, PhD; Dennis Della Corte, Dr.rer.nat

• ***Context.***—**Automated prostate cancer detection using machine learning technology has led to speculation that pathologists will soon be replaced by algorithms. This review covers the development of machine learning algorithms and their reported effectiveness specific to prostate cancer detection and Gleason grading.**

Objective.—**To examine current algorithms regarding their accuracy and classification abilities. We provide a general explanation of the technology and how it is being used in clinical practice. The challenges to the application of machine learning algorithms in clinical practice are also discussed.**

Data Sources.—**The literature for this review was identified and collected using a systematic search. Criteria were established prior to the sorting process to effectively direct the selection of studies. A 4-point system was implemented to rank the papers according to their relevancy. For papers accepted as relevant to our metrics, all cited and citing studies were also reviewed. Studies were then categorized based on whether they implemented binary or multi-class classification methods. Data were extracted from papers that contained accuracy, area under the curve (AUC), or κ values in the context of prostate cancer detection. The results were visually summarized to present accuracy trends between classification abilities.**

Conclusions.—**It is more difficult to achieve high accuracy metrics for multiclassification tasks than for binary tasks. The clinical implementation of an algorithm that can assign a Gleason grade to clinical whole slide images (WSIs) remains elusive. Machine learning technology is currently not able to replace pathologists but can serve as an important safeguard against misdiagnosis.**

(*Arch Pathol Lab Med*. doi: 10.5858/arpa.2022-0460-RA)

The adoption of WSI scanners in clinical practice was accelerated by US Food and Drug Administration approval in 2017, which allowed primary pathologic diagnoses to be made on scanned images. Images in the digital domain allow the application of pathology artificial intelligence (AI), including clinical decision support with algorithms performing specific diagnoses.[1,2] These algorithms, if trained properly, could go beyond the ability of human observation to detect and quantify features that are not recognizable by human perception.[1,3,4]

Prostate cancer is an ideal target for AI diagnostic support. Criteria for prostate adenocarcinoma diagnosis from histologic slides are well defined, and a corresponding grading system known as the Gleason system provides prognostic information and guidelines for treatment.[5] Prostate cancer is the second most frequent cancer and the fifth leading cause of death in men. Pathologists diagnose prostate cancer by examining 6 to 12 needle core biopsies of the prostate. The already-problematic shortage of pathologists is expected to worsen, driving the need for the implementation of AI-based screening tools.[6,7] The Gleason grade system, which denotes the advancement of cancer in the tissue, suffers from diagnostic variation among pathologists. Pathology AI is intended to reduce diagnostic variation and increase the productivity of pathologists, leading to improved patient treatment.[8]

This review summarizes recent literature regarding the use of AI to diagnose prostate adenocarcinoma. Its purpose is to provide an accessible explanation of the current machine learning methods and an overview of the current state of classification ability for working surgical pathologists. It begins by outlining the systematic review process that was followed to produce this literature review. A general machine learning algorithm based on the algorithms uncovered in the literature will then be described, following which the methods used to compare these algorithms will be presented and justified. Finally, a summary of the results found, both qualitative and quantitative, will be given and future implications will be considered.

SYSTEMATIC REVIEW PROCESS

The research process used to conduct this literature review is summarized in Figure 1. The primary questions guiding this review were: What is the current state of AI-driven

Accepted for publication May 4, 2023.

Supplemental digital content is available for this article. See text for hyperlink.

From the Department of Physics and Astronomy, Brigham Young University, Provo, Utah. Frewing and Gibson contributed equally to this work.

The authors have no relevant financial interest in the products or companies described in this article.

Corresponding author: Dennis Della Corte, Dr.rer.nat, Department of Physics and Astronomy, N361 ESC, Brigham Young University, Provo, UT 84602 (email: Dennis.DellaCorte@byu.edu).

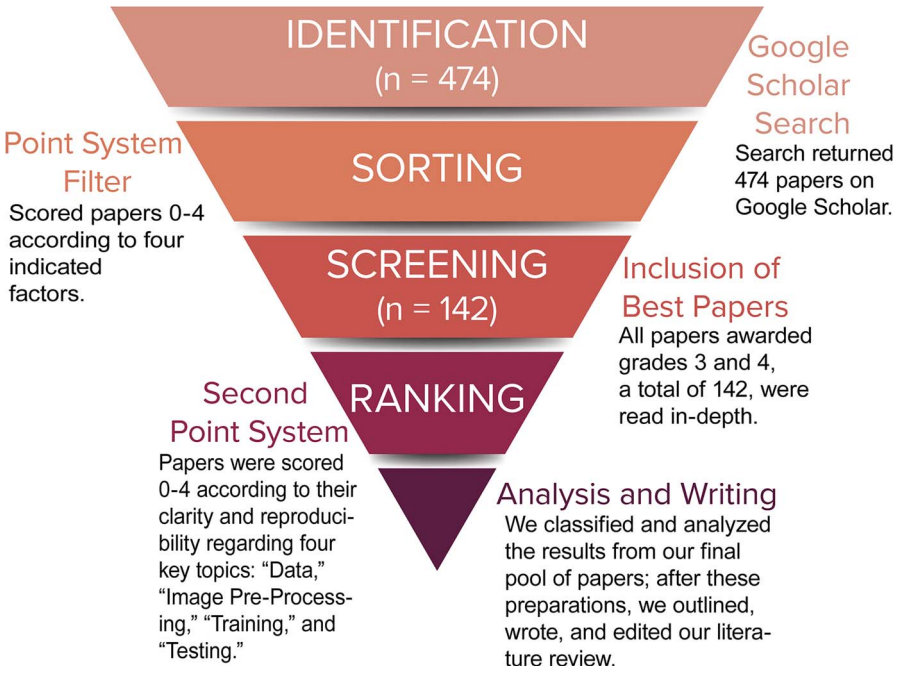

Figure 1. *Process of the systematic literature review. The sorting and screening sections refer to how papers were selected based on their relevancy. The ranking section describes the in-depth review that was performed to guide organization and writing.*

automatic Gleason grading of prostate WSIs? Are any machine learning algorithms currently being implemented in clinical practice, and if not, what are the most pressing and current functional challenges to the implementation of this technology? What are the common grading methods used by machine learning technology?

Search Terms and Criteria

The search expression was broadly descriptive of AI, as there are many terms commonly used to describe machine learning. Although broad, the search expression required that a given paper include the words "Gleason" and "prostate" The search criteria expression was "Gleason" AND ("cancer" OR "adenocarcinoma") AND "prostate" AND ("machine learning" OR "artificial intelligence" OR "neural network" OR "deep learning") AND ("H&E" OR "hematoxylin and eosin") AND ("WSI" OR "whole slide").

Point System Filter

The database selected was Google Scholar; only studies published after 2017 were considered. The initial search was conducted on March 31, 2022; it yielded a total of 474 results. Relevant papers that were published while this literature review was being written were noted and included.

For the filtering process, we created a system based on 4 key factors. The study received 1 point for each factor satisfactorily checked. These factors were established to maximize uniformity and minimize human subjectivity in the sorting and classification processes. The 4 factors were: Does the paper use a machine learning model aimed at detecting prostate cancer? Does the paper contain a quantifiable measurement of their machine learning model's performance? Does the paper compare its performance with that of actual pathologists, or does it consider the model's performance in a real-world clinical setting? Does the paper present a high-impact feature, such as a considerable number (>20) of others who have cited it, or any other novel feature that may make it useful in comparison?

Of the 474 papers we initially reviewed, 40 papers were awarded a score of 4; 78 were awarded a 3; 145 were awarded a 2; and 211 were awarded a 1 or 0.

Forwards/Backwards Search

From this point, we conducted a forward and backward search of the literature that received at least 3 points. For this, all papers cited by a relevant paper and those citing it were manually investigated and considered as potentially relevant. Papers that were added during the forward and backward search were also subjected to the point-system filter; accordingly, 24 papers were added to the existing pool of 118 papers scoring 3 or higher. We ultimately identified a total of 142 studies (see full list in the Supplemental Table in the supplemental digital content) with a score of at least 3 points.

In-Depth Review

To further categorize these 142 studies, we examined each paper in depth according to 4 designated topics. The 4 topics were data sets, preprocessing techniques, training strategies, and testing approach. For each article, the relevant information regarding these topics was extracted and recorded.

MACHINE LEARNING

Introduction

The field of computer-aided diagnosis for prostate cancer relies on the skillful application of deep learning systems. Deep learning systems imitate the human brain to glean patterns from copious amounts of data.[9] When applied to cancer, these systems accurately detect and classify prostate tissue in WSIs.[10] The modern prostate cancer detection algorithm uses a multilayered neural network as its backbone.[11–14] The network of choice is almost exclusively the convolutional neural network (CNN), which mimics the human visual system.[15] The first CNN was introduced in 1989, when LeCun et al[16] created LeNet to recognize handwritten zip code digits in a data set provided by the US Postal Service. Today, dozens of competitive CNNs are

being used to detect cancer. The generic machine learning process that will be explained in this section is summarized in Figure 2.

Data Sets

Regarding data sets, the idiom "what you get out is what you put in" applies quite well. If the data set used to train an algorithm is small and of inferior quality, then the algorithm will produce poor results or have low generalizability to clinical practice.[17–19] Machine learning algorithms must be trained on large, high-quality data sets, which are ideally composed of WSIs representative of those seen in clinical practice.[20–24] *Quality* as used here refers to the resolution of the WSI scan and the kind of annotation used to label it. The annotation of the WSI in most data sets is at least overseen by a pathologist and can range from a single slide–level label of an overall Gleason score (GS) or Gleason grade[25] to individual gland–level Gleason pattern assignments.[26,27] Some studies attempt to strengthen the generalizability of their algorithm by using multiple data sets because annotation variation exists in clinical practice.[26,28–33]

Another point that must be considered is that annotations are only as good as the person making them. Observer variability is known to be an issue to the consistency of Gleason grading among pathologists,[34] leading some researchers to look at possible genetic indicators for cancer detection instead of hematoxylin and eosin (H&E) slide annotation.[35] There are a number of factors that could explain the difference in Gleason grading among pathologists, including the experience of the pathologist assigning the grade, the time spent examining the slide, or the quality of the image. Assigning a Gleason grade is somewhat subjective to the pathologist. Training an algorithm on data that may be specific to an individual or a small group of pathologists can limit its generalizability to clinical practice.[27] One solution that has been implemented to enhance generalizability is a model wherein multiple pathologists' annotations are merged and averaged into a single ground truth label for training purposes.[36]

Pretraining/Transfer Learning

Just as humans learn patterns in one domain and apply those patterns elsewhere, strategically designed algorithms can also transfer what they learn. The modern prostate cancer detection algorithm first learns patterns from a generic data set. Popular generic data sets used in the literature are ImageNet, Microsoft Common Objects in Context, and Canadian Institute for Advanced Research, all of which include millions of annotated images with hundreds of classes.[37,38] A model trained on a generic data set learns general image detection principles, such as the detection of edges, shapes, and objects; this is an algorithm's version of first learning to walk before learning to run.

A pretrained model has a head start when applied to the image detection task of classifying prostate tissue WSIs.[39] The already-experienced model is better at detecting cancer than a nascent model, which has not learned general patterns from pretraining. Today's cancer detection machine learning models often learn patterns from one type of cancer and apply these patterns to another type of cancer. This method of machine learning, known as cross-domain transfer learning, is making cancer detection more accurate and bridges the gaps between models trained on different types of tissue. For example, a model that pretrains on breast cancer WSIs and applies its knowledge to prostate cancer WSIs is far more accurate than a model that learns without pretraining, beginning training from a blank slate.[29,40]

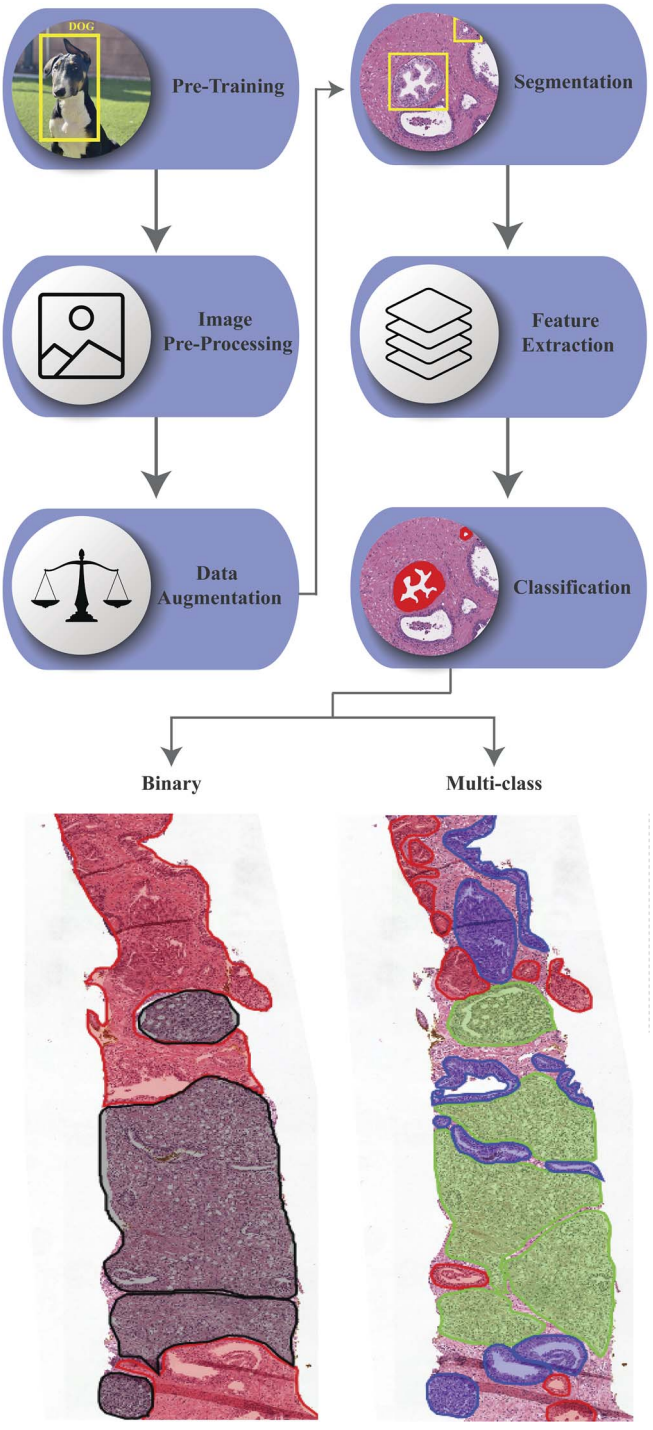

Figure 2. *General representation of the process by which a neural network is trained to differentiate whole slide images. At the end of the flow chart the line is broken into 4 categories; for detailed discussion see the Methods section.*

Image Preprocessing

Image preprocessing prepares images in a data set for use in training. It enhances the contents of a data set to improve the performance of the overall machine learning model.[41–43]

In this process the overall qualities of each image, such as color, orientation, and size, are normalized.[27] Hospitals and medical centers commonly have unique staining protocols and varying whole slide scanners, so prostate tissue data sets suffer from intercenter variation in magnification, stain color saturation, image noise, etc.[44–46] To combat this variation, a cancer detection algorithm uses a preprocessing step to make a data set's images ideally formatted for the machine learning algorithm.[47–53]

Generally, a cancer detection algorithm would perform 3 main functions during image preprocessing: filtering (or tiling), color normalization, and image denoising. A generic image preprocessing step would proceed as follows:

First, a sliding box, termed "filter," is passed over a WSI at different magnifications.[28] Each frame of reference produced as the box slides over the WSI creates a unique subimage, called a *tile*.[17,29,54] Therefore, a region of tissue may exist in multiple tiles if the tiles are of different sizes.[55] This allows the machine learning model to learn from multiple magnification levels of the same tissue.

Next, each tile is color normalized to produce a set of images with homogeneous stain distribution.[29,56] If 2 tiles have significantly different saturations, then they are transformed into tiles of similar stain profiles.[57]

Third, the tiles are denoised.[58] In a data set, image noise exists as unwanted, random brightness or color variation produced by a scanner's image sensor. Noise, which negatively affects an algorithm's performance, can be reduced before each tile is passed onward in the network.[27,46] It is clear that images in a data set can be preprocessed to increase the performance of a model; a model's accuracy can also be further improved by augmenting the resulting tiles from the preprocessing stage.[59]

Data Augmentation

An algorithm is not limited in learning to the tiles from the image preprocessing phase. In fact, most algorithms augment these tiles to produce a more robust data set. A model's performance is dependent on the size of the data set it is trained on.[21,23] An algorithm can leverage this principle by artificially increasing, or augmenting, the size of a data set.[24] It is also necessary to augment the distribution of tissue classes, so that a data set is balanced for each class of tissue. The most straightforward augmentations are achieved by transforming tiles.

The most common tile transformations used to augment data sets include mirroring, rotating, differentiable zooming, and scaling tiles.[60–62] Class balancing is commonly achieved by oversampling underrepresented classes (reusing the same data points repeatedly) or undersampling overrepresented classes (selecting a random subsample of data points).

The tile transformations and class distribution enhancements increase model performance; thus, the best-performing prostate cancer detection algorithms use both types of augmentations.[63,64]

Segmentation

After preprocessing and augmenting, the machine learning model looks at each tile with a granular approach.[65,66] As pixel-wise occurrences like artifacts, cribriform patterns, or Gleason patterns are detected in each tile, the network associates related pixels with a certain class.[67,68] When 2 separate instances of the same tissue differentiation occur, the algorithm classifies them as similar but unique cases.[69] A region of tissue can be segmented multiple times—once for each magnification level. Segmentation for prostate cancer detection and classification produces *regions of interest* (ROIs), which are areas of notable tissue differentiation.[30,70–73] The ROIs at different magnifications are passed onward so the model can extract features from them; the uninteresting, nondifferentiated tissue is excluded.[74–76]

Feature Extraction

During feature extraction, a network detects features in ROIs to simplify complex data. Features are numbers that represent specific qualities of the ROI.[77] Prostate cancer detection algorithms extract features that describe the concentration of nuclei, the sizes of tumors, and even the borders of malignant tissue.[78,79] With tiles and ROIs of the same tissue area present in different magnifications, the features must be combined in a meaningful way.[29,65] The tiles' data are merged to decrease computational expenses while maintaining algorithm performance; the tiles of prostate tissue are represented by a grid filled with numbers, also known as a *matrix*. The matrices are stacked to form a three-dimensional matrix stack that contains the feature information of the preliminary matrices.[1,10] The higher priority the feature, the more weight the feature holds in the new matrix.[80,81] After the data are stacked in a process called *pooling*, the feature information is used to delegate ROIs into specific classes during the classification step.[82,83]

Classification

Classification is the culmination of a prostate cancer detection algorithm.[84] All previous processes, transformations, and computations lead to the final class predictions of a model. These predictions are made by a fully convolutional (FC) head. The FC head is the last step, or *layer*, in a neural network. The flattened data from the previous pooling layer are given to the FC head, which ultimately decides how an ROI should be classified according to the probability that the region belongs to a certain class.[28,85] Classification methods can be binary (between 2 outputs) or multi-class (among more than 2 outputs); the FC head's classification method is determined before the neural network is trained.[86–91]

METHODS

Classification Groups

After the in-depth review, we performed an aggregation of the reported classification methods. Studies with explicit numerical results were examined and categorized according to 4 classification categories, which are depicted in Figure 3.[74] We found 65 papers containing results useful for comparison. These categories reflect 4 general types of classification approaches taken in the literature reviewed.

The first group, labeled Binary 1, contains studies that implemented machine learning methods to distinguish between cancerous and noncancerous regions of the prostate WSI. Also included in this category were studies that labeled regions of the WSI as benign or malignant and those that labeled slides as suspicious or nonsuspicious. Some applied a binary label to the entire WSI, whereas others applied the label to specific regions on the slide. This category is designated as Binary 1 because the decisions made by the machine learning models were binary in nature: suspicious/nonsuspicious, benign/malignant, cancerous/noncancerous.

The second group, titled Binary 2, refers to studies in which machine learning models made binary decisions but in the context of a Gleason classification. Included in Binary 2 are those models that could distinguish between WSIs with a GS of greater than 6 or

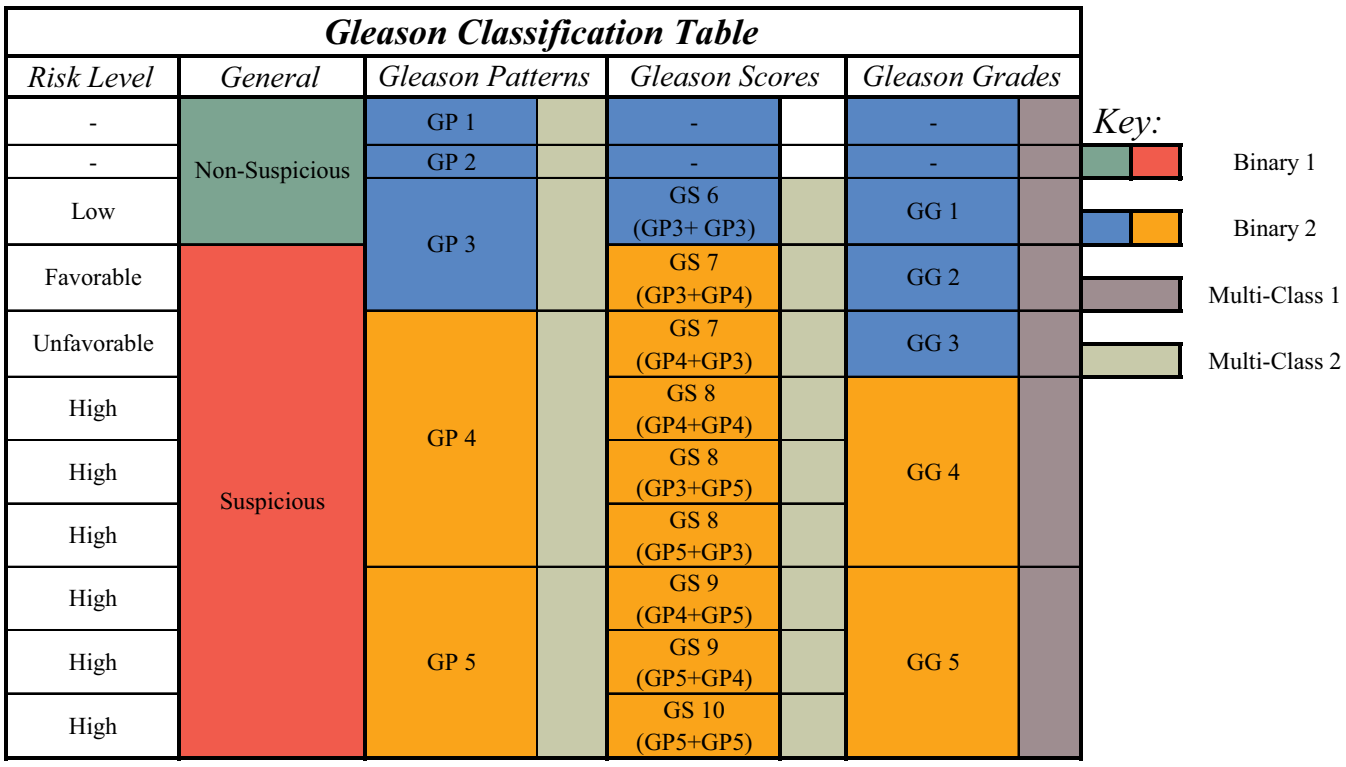

Figure 3. *Overview of risk levels associated with common whole slide image labels. Color key maps these labels against 4 classes of algorithms found in the literature (see Methods section for definition of these classes). Abbreviations: GG, Gleason Grade Group; GP, Gleason pattern; GS, Gleason score.*

less than or equal to 6 (GS >6 versus GS ≤6). Some distinguished between slides with a Gleason Grade Group of 3 or greater than 3, and others distinguished between a Gleason pattern of 3 and a Gleason pattern of 4 or greater. As seen in the chart, these distinctions are similar in their nature and difficulty; thus, they were combined into one category.

The third group, called Multiclassification 1, contains studies that attempted to perform more than one classification, specifically in terms of Gleason Grade Groups. Studies included in this category distinguish Gleason Grade Groups 0 through 5, with Gleason Grade Group 0 in this context referring to benign tissue. Studies that distinguished between Gleason Grade Groups 2 through 5 or 3 through 5 were also included in this category. These studies are distinct from those in Binary 2, which classified only between slides that had a Gleason Grade Group of 3 or greater than 3. Although the distinction between Gleason Grade Groups 3 and 4 is critical, the distinction between Gleason Grade Groups 4 and 5 also affects treatment plans and the prospects of surgery.

Finally, the fourth group, which is called Multiclassification 2, refers to studies that performed more than one classification but not specifically in terms of a Gleason Grade Group. These studies classified Gleason patterns 1 through 5 or 3 through 5. Some studies classified their slides as Gleason score 6 through 10.

Measurements Taken Into Consideration

The κ values, AUC values, and accuracy values were collected for each of the 4 classification groups. All 3 values were collected and included in Figure 4 if they were reported. For example, if both a κ and an AUC value were reported, both values were added to the figure. This could raise the concern of counting the same paper twice, but by separating these values into distinct categories, the additional data points facilitate comparison among different accuracy measurements. Sometimes algorithms were tested on multiple data sets or trials and thus had multiple reported values for the same measurement. In these cases, 1 of each given measurement, either accuracy, κ, or an AUC value, was selected. Some studies included additional measurements like F-1 scores and sensitivity and specificity values to quantify accuracy, but they were excluded, as infrequent use prevented comparison.

The accuracy values are determined by comparing the number of correct responses and the number of incorrect responses given by the machine learning algorithm. Accuracy for machine learning models can be calculated using Equation 1, where TP, TN, FP, and FN refer to true-positive, true-negative, false-positive, and false-negative, respectively.[92]

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \quad (1)$$

In many cases the exact methodology for how the accuracy value was calculated was not reported. In some cases, the accuracy value was a mean taken over several trials performed with the machine learning model.[93]

The AUC value balances the accuracy of the true-positive and false-positive values. A model that got 100% of the tests wrong would have an AUC value of 0, and inversely, if 100% of the tests were right it would have an AUC value of 1. The AUC value is quite literally the area under a curve on a graph whose x-axis is the false-positive value on a scale of 0 to 1 and whose y-axis is the true-positive value on a scale of 0 to 1.[69,89,94,95] If the AUC achieves a value of 1, this means it was 100% accurate in its tasks.[39] AUC values are often used to quantify accuracy, but, as can be observed in the scatterplot, there is little variability in the individual values among studies. Thus, the AUC value is a poor metric with which to quantify accuracy because it does very little to distinguish among developed algorithms.[96] A better accuracy metric for comparison is the κ value.

The κ value is a complex statistical measurement that considers several factors including the number of classification groups, the number of images classified, the number of raters and the expected

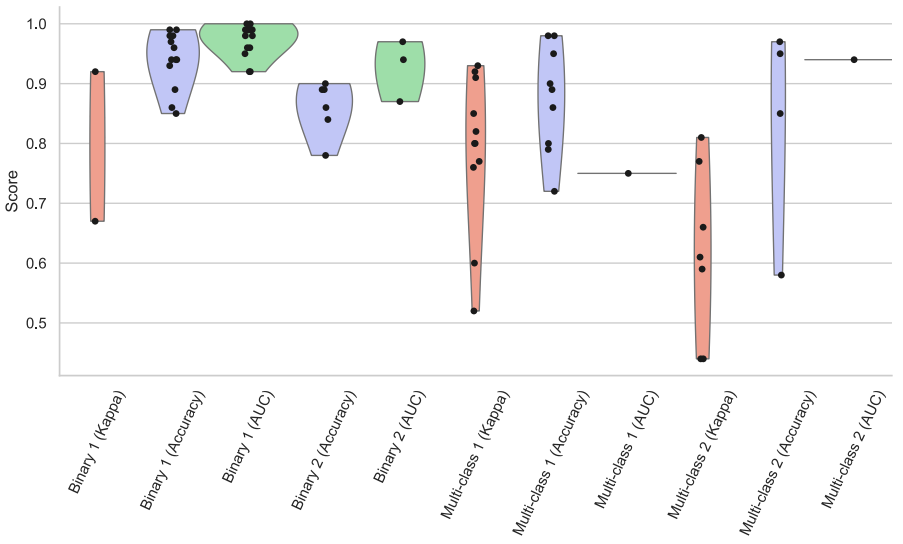

Figure 4. *Scatterplot of area under the curve (AUC) values, accuracy values, and κ values for the studies examined. As machine learning algorithms progress from binary classification to multiclassification efforts, a greater variability can be observed in the results reported.*

values of those ratings.[36] The κ value is defined by Equation 2, where P_o refers to the observed agreement among the raters (ie, the pathologists) and P_e is the hypothetical probability of chance agreement. K is the κ value.

$$K = \frac{(P_o - P_e)}{(1 - P_e)}. \quad (2)$$

If the raters have perfect agreement, the κ value is 1, and if there is nothing more than random agreement among raters, the κ value is 0.[92] In some situations, a weighted κ value was used, which ensures that highly erroneous classifications are penalized heavily.[97] The κ value helps balance observer variability and is commonly used in this domain of machine learning.[98] To simplify, the κ value helps to quantify the concordance between a defined ground truth and the machine learning algorithms' classification.[30,99] It is also used to compare pathologists' scores with each other to quantify differences in individual ratings.[58]

RESULTS

The quantitative results of this literature review are summarized in Figure 4, which depicts the accuracy values across the 4 categories of classification methods. From left to right (binary to multiclassification) in the scatterplot a general trend can be seen downwards for accuracy values, and a trend for increased variability. This shows that as algorithms move beyond binary classification it is more difficult to achieve high accuracy values.[80] It is also important to note that the algorithms that performed a binary classification task were in much better agreement with experts. These results suggest that binary classification of prostate tissue from H&E WSIs is in many cases accurate and reliable.

The increased variability seen in the multiclassification tasks can be explained by the increasing difficulty of the task and by the type of statistical measurement used to quantify accuracy. Most studies that performed binary classification were able to report a nearly perfect AUC value, which suggests the AUC value is not ideal for comparison among algorithms. There is much less variation among AUC values in comparison with κ values, which were more commonly used to quantify accuracy among studies that performed multiclassification tasks. The κ values incorporate more information into the statistical measure, which allows for better comparison of performance among algorithms. Overall, the graph shows the increasing difficulty of multiclassification tasks and presents a general overview of the current state of the literature with regard to prostate cancer detection.

The qualitative results of this literature review were extracted and compiled by an in-depth review. The clinical implementation of this technology remains minimal in nature and has been hindered by observer variability.

DISCUSSION

Observer variability is a challenge that must be taken seriously in the development of machine learning algorithms.[34,100] The fact that Gleason grading is somewhat subjective based on the grader means that it can be difficult to assign a "correct" grade to every image for a machine learning model.[101] In a case study, an expert team of 4 pathologists, including a genitourinary specialist, labeled a set of 331 slides with Gleason gradings. Afterwards, 29 pathologists were tasked to assign individual Gleason grades to the same slides. Upon comparison, an average Gleason grading accuracy value of 0.61 was found between the expert team and individual pathologists, with individual accuracies ranging between 0.31 and 0.74.[102] Another study found the agreement among 24 pathologists in the context of Gleason grading to be a κ value of 0.67.[98]

Given the discrepancy among grades assigned by pathologists, it is difficult to define a minimum "pathologist accuracy" threshold a machine learning algorithm needs to pass. Comparing a machine learning output directly against one specific pathologist is a poor performance indicator because of different experience levels and specializations. There is a significant amount of variation among pathologists when it comes to Gleason grading of WSIs.[103] This variation makes it difficult to train an algorithm, because the annotations of a single pathologist do not establish an absolute truth in terms of Gleason grading. The existing variance in label quality suggests that a variety of annotations from different pathologists should be taken to create a better ground truth in terms of Gleason grading. Instead of just assigning a single value, the AI should instead also report a confidence score that could resemble the likely discrepancy among an expert group, potentially encoded as grade variance. As it stands, training an algorithm that classifies Gleason grades with 100% accuracy is impossible because "accuracy" here is the interpretive consensus of a group of pathologists. Instead of an expert

panel consensus, recording the distribution of Gleason grades assigned by individual pathologists would yield a better label set. In this situation an AI would be trained to predict an expected distribution of expert opinions, rather than that of a single expert. We think that such a model would be far better at supporting pathologists, as it would de facto represent a crowd response.

In the absence of such a model, it may be best to view the clinical implementation of machine learning algorithms as a way to focus pathologists on the most relevant tissue areas, to help reduce variability among pathologists, and to provide a second opinion at little cost, which would enhance care and treatment options for patients[104,105] and serve as a safety check against any misdiagnosis.[106–109]

As expected, multiclassification tasks are very difficult for current AI models. As mentioned previously and highlighted in Figure 4, there is a noticeable downward trend in accuracy from the binary classification to the multiclassification tasks. Unsurprisingly, the increasing number of categories deteriorates precision and accuracy because each additional category represents an increasingly smaller slice of all the possible outcomes. The lower accuracy of multiclassification tasks is also related to the subjectivity inherent in the Gleason grading system—Gleason Grade Groups are not based on mathematical measurements like tumor length, quantity, or circumference. Although these could be taken into consideration by a pathologist, there is no set mathematical scale by which Gleason Grade Groups are classified. They are roughly based on the morphologic features present, but again, this is subjective to the pathologist or experts examining the H&E stain. Even among experts there is considerable variation in classification of WSIs.[103]

As Gleason grading is a difficult subjective measure, other features taken into consideration by pathologists may present attractive alternative objectives for AI to learn. Such features include the quantification of tumors, perineural invasion, or intraductal and cribriform pattern. Perineural invasion describes the invasion of cancer in the immediate periphery of nerves. Nerves can serve as a route for metastatic spread, and thus an invasion can predict irreversible metastasis. Some algorithms were trained to recognize this element in addition to other factors with marked success.[110] Intraductal carcinoma, a proliferation of cancer in prostate ducts, can also be a training objective. Some algorithms take this morphologic feature into consideration when grading a WSI.[107] Cribriform pattern on a tumor is associated with carcinoma and can serve as an indicator for areas of the prostate biopsy that need additional attention. There are several developed algorithms that can detect the presence of cribriform pattern.[77,78]

The presence of cribriform pattern, perineural invasion, intraductal carcinoma, or other details such as tumor length and density can factor into an expert's diagnosis and treatment plan. Future AI algorithms would be well advised to consider these features, as their prediction could aid pathologists greatly in their efficiency and decision making.

For those who worry about replacing pathologists by AI algorithms, comfort can be found in recent examples of parallel AI applications failing to adapt to the level of variation that exists in the real world. IBM attempted to develop an AI doctor called Watson back in 2011, but after investing billions into the project, it was scrapped entirely by 2014.[111] Although AI may one day be capable of replacing doctors or pathologists, that day has not yet arrived. On that day—if AI ever achieves general human-level intelligence—pathology will be just one among countless professions threatened by AI. Until that day, refusing to use AI as a tool for augmenting a pathologist may be compared with stubbornly refusing email in favor of paper mail. Although arguments can be made, efficiency will likely favor willing pathologists that thoroughly vet and carefully choose suitable AI to amplify their output. Based on current trends, AI will and most likely be limited to serve as an assistant for pathologists in the foreseeable future.

CONCLUSIONS

Clinical Implementation

In conclusion, machine learning algorithms will not replace human pathologists in the near future, but instead may offer a useful tool to help decrease the work burden and increase the accuracy of practicing pathologists.[69,80,84,112–114] One of the most well-known tools being implemented to date is Paige Prostate (Paige AI).[115,116] Currently the Paige Prostate software is able to detect and label carcinoma as well as provide a Gleason score and other quantifiable measurements of the tumors. Paige Prostate has also demonstrated its ability to decrease time spent by pathologists examining slides and increase overall accuracy of assigned Gleason grades.[99] One important consideration is that for time to be saved by the pathologist, there must be a level of trust in the algorithm. If there is little trust, then no time will be saved, because the pathologist may be spending time double-checking the algorithm.

Others found comparable results regarding accuracy improvement when machine learning methods were used.[34,58] One specific tool, developed by researchers at the University of Wisconsin, was also implemented to improve the accuracy of 3 pathologists ($\kappa = 0.56–0.70$ to $\kappa = 0.88–0.93$).[58] Another tool called Galen Prostate (Ibex-AI) has been in use in Israel as a quality control for pathologists.[110] Two other prostate screening tools recognized in the literature and cleared for clinical implementation by a governing regulatory body are Deep-Dx[117] (DeepBio) and Inify[94] (Inify Laboratories).

The expanding field of AI-integrated pathology presents promising possibilities.[118] This innovative technology may decrease workloads and increase accuracy for practicing pathologists.[116] One hurdle to clinical implementation, which may be better viewed as an opportunity, is the widespread adoption of whole slide scanners. Effective implementation of this technology in clinical practice necessitates the adoption of whole slide scanners. High-quality digital slides that can be uploaded and evaluated easily by pathologists are necessary for machine learning algorithms as well.[21,119] Looking beyond the United States and Europe, the ability to upload slides and view them digitally from afar can help improve access to health care in the developing world where pathologists are scarce.[22,120]

Limitations of This Study

This study is limited to the results given in the initial Google search, or which were derived from a forward and backward search. We also limited our literature to studies published between 2018 and late 2022. Furthermore, some aspects of the sorting process were inherently subjective to the research assistant whose responsibility it was to review the literature and categorize it. Some generalizations were also made to present the 4 categories of classification.

We thank TJ Hart for his contributions in identifying relevant algorithms. We thank Brigham Young University for providing undergraduate research funds.

References

1. Cimadamore A, Cheng L, Scarpelli M, Lopez-Beltran A, Montironi R. Digital diagnostics and artificial intelligence in prostate cancer treatment in 5 years from now. *Transl Androl Urol*. 2021;10(3):1499–1505. doi:10.21037/tau-2021-01

2. Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun*. 2021;41(11):1100–1115. doi:10.1002/cac2.12215

3. Gupta R, Kurc T, Sharma A, Almeida JS, Saltz J. The emergence of pathomics. *Curr Pathobiol Rep*. 2019;7(3):73–84. doi:10.1007/s40139-019-00200-x

4. Durkee MS, Abraham R, Clark MR, Giger ML. Artificial intelligence and cellular segmentation in tissue microscopy images. *Am J Pathol*. 2021;191(10):1693–1701. doi:10.1016/j.ajpath.2021.05.022

5. Chen N, Zhou Q. The evolving Gleason grading system. *Chin J Cancer Res*. 2016;28(1):58–64. doi:10.3978/j.issn.1000-9604.2016.02.04

6. Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *J Pathol*. 2019;249(2):143–150. doi:10.1002/path.5310

7. Gross DJ, Robboy SJ, Cohen MB, et al. Strong job market for pathologists results from the 2021 College of American Pathologists Practice Leader Survey. *Arch Pathol Lab Med*. 2023;147(4):434–441. doi:10.5858/arpa.2022-0023-CP

8. He Y, Zhao H, Wong ST. Deep learning powers cancer diagnosis in digital pathology. *Comput Med Imaging Graph*. 2021;88:101820. doi:10.1016/j.compmedimag.2020.101820

9. Otálora S, Atzori M, Khan A, Jimenez-del-Toro O, Andrearczyk V, Müller H. Systematic comparison of deep learning strategies for weakly supervised Gleason grading. *Proc SPIE Med Imaging Digit Pathol*. 2020;11320:142–149. doi:10.1117/12.2548571

10. Ikromjanov K, Bhattacharjee S, Hwang Y-B, Sumon RI, Kim H-C, Choi H-K. Whole slide image analysis and detection of prostate cancer using vision transformers. *2022 Intl Conf Artif Intelligence Inf Comm*. 2022:399–402. doi:10.1109/ICAIIC54071.2022.9722635

11. Pohjonen J, Stürenberg C, Rannikko A, Mirtti T, Pitkänen E. Spectral decoupling for training transferable neural networks in medical imaging. *iScience*. 2022;25(2):103767. doi:10.1016/j.isci.2022.103767

12. Yang B, Xiao Z. A multi-channel and multi-spatial attention convolutional neural network for prostate cancer ISUP grading. *Appl Sci*. 2021;11(10):4321. doi:10.3390/app11104321

13. Duran-Lopez L, Dominguez-Morales JP, Rios-Navarro A, et al. Performance evaluation of deep learning-based prostate cancer screening methods in histopathological images: measuring the impact of the model's complexity on its processing speed. *Sensors*. 2021;21(4):1122. doi:10.3390/s21041122

14. Aryal M, Soltani NY. Context-aware graph-based self-supervised learning of whole slide images. *IEEE Intl Conf Acoustics Speech Signal Processing*. 2022:3553–3557. doi:10.1109/ICASSP43922.2022.9747899

15. Ström P, Kartasalo K, Olsson H, et al. Pathologist-level grading of prostate biopsies with artificial intelligence [published online July 2, 2019]. *arXiv*. doi:10.48550/arXiv.1907.01368

16. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput*. 1989;1(4):541–551. doi:10.1162/neco.1989.1.4.541

17. Komura D, Ishikawa S. Advanced deep learning applications in diagnostic pathology. *Transl Regul Sci*. 2021;3(2):36–42. doi:10.33611/trs.2021-005

18. Otálora S, Atzori M, Andrearczyk V, Khan A, Müller H. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Front Bioeng Biotechnol*. 2019;7:198. doi:10.3389/fbioe.2019.00198

19. Bhattacharjee S, Ikromjanov K, Carole KS, et al. Cluster analysis of cell nuclei in H&E-stained histological sections of prostate cancer and classification based on traditional and modern artificial intelligence techniques. *Diagnostics*. 2021;12(1):15. doi:10.3390/diagnostics12010015

20. Ayyad SM, Shehata M, Shalaby A, et al. Role of AI and histopathological images in detecting prostate cancer: a survey. *Sensors*. 2021;21(8):2586. doi:10.3390/s21082586

21. 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. doi:10.1038/s41591-019-0508-1

22. Han W, Johnson C, Gaed M, et al. Automatic cancer detection and localization on prostatectomy histopathology images. *Proc SPIE Med Imaging Digit Pathol*. 2018;11320:205–212. doi:10.1117/12.2292450

23. Luca AR, Ursuleanu TF, Gheorghe L, et al. Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. *Inform Med Unlocked*. 2022;29:100911. doi:10.1016/j.imu.2022.100911

24. 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*. 2019;24(5):1413–1426. doi:10.1109/JBHI.2019.2944643

25. Pinckaers H, Bulten W, van der Laak J, Litjens G. Detection of prostate cancer in whole- slide images through end-to-end training with image-level labels. *IEEE Trans Med Imaging*. 2021;40(7):1817–1826. doi:10.1109/TMI.2021.3066295

26. Otálora S, Marini N, Müller H, Atzori M. Combining weakly and strongly supervised learning improves strong supervision in Gleason pattern classification. *BMC Med*. 2021;21(1):77. doi:10.1186/s12880-021-00609-0

27. Linkon AHM, Labib MM, Hasan T, Hossain M. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: an extensive study. *Inform Med*. 2021;24:100582. doi:10.1016/j.imu.2021.100582

28. Homeyer A, Geißler C, Schwen LO, et al. Recommendations on test datasets for evaluating AI solutions in pathology [published online April 22, 2022]. *arXiv*. doi:10.48550/arXiv.2204.14226

29. Schaer R, Otálora S, Jimenez-del-Toro O, Atzori M, Müller H. Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. *J Pathol Inform*. 2019;10(1):19. doi:10.4103/jpi.jpi_88_18

30. Schmidt A, Silva-Rodríguez J, Molina R, Naranjo V. Efficient cancer classification by coupling semi supervised and multiple instance learning. *IEEE Access*. 2022;10:9763–9773. doi:10.1109/ACCESS.2022.3143345

31. Marini N, Otálora S, Müller H, Atzori M. Semi-supervised learning with a teacher-student paradigm for histopathology classification: a resource to face data heterogeneity and lack of local annotations. *Proc ICPR Intl Workshops Challenges*. 2021:105–119. doi:10.1007/978-3-030-68763-2_9

32. Arvaniti E, Claassen M. Coupling weak and strong supervision for classification of prostate cancer histopathology images [published online November 16, 2018]. *arXiv*. doi:10.48550/arXiv.1811.07013

33. Marini N, Otálora S, Müller H, Atzori M. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: an experiment on prostate histopathology image classification. *Med Image Anal*. 2021;73:102165. doi:10.1016/j.media.2021.102165

34. 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. doi:10.1016/S1470-2045(19)30739-9

35. Dadhania V, Gonzalez D, Yousif M, et al. Leveraging artificial intelligence to predict ERG gene fusion status in prostate cancer. *BMC Cancer*. 2022;22(1):494. doi:10.1186/s12885-022-09559-4

36. Qiu Y, Hu Y, Kong P, et al. Automatic prostate Gleason grading using pyramid semantic parsing network in digital histopathology. *Front Oncol*. 2022;12:772403. doi:10.3389/fonc.2022.772403

37. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov*. 2021;11(4):900–915. doi:10.1158/2159-8290.CD-21-0090

38. Silva-Rodríguez J, Colomer A, Naranjo V. WeGleNet: a weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. *Comput Med Imaging Graph*. 2021;88:101846. doi:10.1016/j.compmedimag.2020.101846

39. Mun Y, Paik I, Shin S-J, Kwak T-Y, Chang H. Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning. *NPJ Digit Med*. 2021;4(1):99. doi:10.1038/s41746-021-00469-6

40. Ikromjanov K, Bhattacharjee S, Hwang Y-B, Kim H-C, Choi H-K. Multi-class classification of histopathology images using fine-tuning techniques of transfer learning. *J Korea Multimed Soc*. 2021;24(7):849–859. doi:10.9717/kmms.2021.24.7.849

41. Salvi M, Molinari F, Acharya UR, Molinaro L, Meiburger KM. Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification. *Comput Methods Programs Biomed Update*. 2021;1:100004. doi:10.1016/j.cmpbup.2021.100004

42. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58:101544. doi:10.1016/j.media.2019.101544

43. Salvi M, Acharya UR, Molinari F, Meiburger KM. The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. *Comput Biol Med*. 2021;128:104129. doi:10.1016/j.compbiomed.2020.104129

44. Duran-Lopez L, Dominguez-Morales JP, Conde-Martin AF, Vicente-Diaz S, Linares- Barranco A. PROMETEO: a CNN-based computer-aided diagnosis system for WSI prostate cancer detection. *IEEE Access*. 2020;8:128613–128628. doi:10.1109/ACCESS.2020.3008868

45. Rana A, Lowe A, Lithgow M, et al. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA Netw Open*. 2020;3(5):e205111. doi:10.1001/jamanetworkopen.2020.5111

46. Swiderska-Chadaj Z, de Bel T, Blanchet L, et al. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Sci Rep*. 2020;10(1):14398. doi:10.1038/s41598-020-71420-0

47. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J*. 2018;16:34–42. doi:10.1016/j.csbj.2018.01.001

48. Duran-Lopez L, Dominguez-Morales JP, Gutierrez-Galan D, et al. Wide & deep neural network model for patch aggregation in CNN-based prostate cancer detection systems. *Comput Biol Med*. 2021;136:104743. doi:10.1016/j.compbiomed.2021.104743

49. Marini N, Atzori M, Otálora S, Marchand-Maillet S, Müller H. H&E-adversarial network: a convolutional neural network to learn stain-invariant

features through hematoxylin & eosin regression [published online January 17, 2022]. *arXiv*. doi:10.48550/ARXIV.2201.06329

50. Ren J, Hacihaliloglu I, Singer EA, Foran DJ, Qi X. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. *Proc Med Imaging Comput Comp Asstc Intervention*. 2018;11071:201–209. doi:10.1007/978-3-030-00934-2_23

51. Rana A, Lowe A, Lithgow M, et al. High accuracy tumor diagnoses and benchmarking of hematoxylin and eosin stained prostate core biopsy images generated by explainable deep neural networks [published online August 2, 2019]. *arXiv*. doi:10.48550/arXiv.1908.01593

52. Khan A, Atzori M, Otálora S, Andrearczyk V, Müller H. Generalizing convolution neural networks on stain color heterogeneous data for computational pathology. *Proc SPIE Med Imaging Digit Pathol*. 2020;11320:173–186. doi:10.1117/12.2549718

53. Anghel A, Stanisavljevic M, Andani S, et al. A high-performance system for robust stain normalization of whole-slide images in histopathology. *Front Med*. 2019;6:193. doi:10.3389/fmed.2019.00193

54. Myronenko A, Xu Z, Yang D, Roth HR, Xu D. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. *Proc Med Imaging Comput Comp Asstc Intervention*. 2021;12908:329–338. doi:10.1007/978-3-030-87237-3_32

55. 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):347–351. doi:10.1016/j.euf.2019.11.003

56. García G, Colomer A, Naranjo V. First-stage prostate cancer identification on histopathological images: hand-driven versus automatic learning. *Entropy*. 2019;21(4):356. doi:10.3390/e21040356

57. Bulten W, Pinckaers H, van Boven H, et al. Automated Gleason grading of prostate biopsies using deep learning [published online July 18, 2019]. *arXiv*. doi:10.48550/arXiv.1907.07980

58. Huang W, Randhawa R, Jain P, et al. Development and validation of an artificial intelligence–powered platform for prostate cancer grading and novel quantification. *JAMA Netw Open*. 2021;4(11):e2132554. doi:10.1001/jamanetworkopen.2021.32554

59. Xu H, Park S, Hwang TH. Computerized classification of prostate cancer Gleason scores from whole slide images. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;17(6):1871–1882. doi:10.1109/TCBB.2019.2941195

60. 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. doi:10.1001/jamanetworkopen.2019.0442

61. Thandiackal K, Chen B, Pati P, et al. Differentiable zooming for multiple instance learning on whole-slide images [published online April 26, 2022]. *arXiv*. doi:10.48550/arXiv.2204.12454

62. Haghighat M, Browning L, Sirinukunwattana K, et al. Automated quality assessment of large digitised histology cohorts by artificial intelligence. *Sci Rep*. 2022;12(1):5002. doi:10.1038/s41598-022-08351-5

63. Eminaga O, Abbas M, Kunder C, et al. Plexus convolutional neural network (PlexusNet): a novel neural network architecture for histologic image analysis [published online August 24, 2019]. *arXiv*. doi:10.48550/arXiv.1908.09067

64. Koziarski M, Cyganek B, Olborski B, et al. DiagSet: a dataset for prostate cancer histopathological image classification [published online May 9, 2021]. *arXiv*. doi:10.48550/arXiv.2105.04014

65. Pérez-Bueno F, Serra JG, Vega M, Mateos J, Molina R, Katsaggelos AK. Bayesian K- SVD for H and E blind color deconvolution: applications to stain normalization, data augmentation and cancer classification. *Comput Med Imaging Graph*. 2022;97:102048. doi:10.1016/j.compmedimag.2022.102048

66. Salvi M, Bosco M, Molinaro L, et al. A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artif Intell Med*. 2021;115:102076. doi:10.1016/j.artmed.2021.102076

67. Patil A, Talha M, Bhatia A, et al. Fast, self supervised, fully convolutional color normalization of H&E stained images. *Proc 2021 IEEE 18th Intl Symp Biomed Imaging*. 2021:1563–1567. doi:10.1109/ISBI48211.2021.9434121

68. Singhal N, Soni S, Bonthu S, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Sci Rep*. 2022;12(1):3383. doi:10.1038/s41598-022-07217-0

69. Oner MU, Ng MY, Giron DM, et al. An AI-assisted tool for efficient prostate cancer diagnosis. *bioRxiv*. Preprint posted online February 9, 2022. doi:10.1101/2022.02.06.479283

70. Peyret R, Khelifi F, Al-Ghreimil N, Al-Baity H, Bouridane A. Convolutional neural network–based automatic classification of colorectal and prostate tumor biopsies using multispectral imagery: system development study. *JMIR Bioinf Biotechnol*. 2022;3(1):e27394. doi:10.2196/27394

71. Li W, Li J, Sarma KV, et al. Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imaging*. 2018;38(4):945–954. doi:10.1109/TMI.2018.2875868

72. Bhattacharjee S, Park H-G, Kim C-H, et al. Quantitative analysis of benign and malignant tumors in histopathology: predicting prostate cancer grading using SVM. *Appl Sci*. 2019;9(15):2969. doi:10.3390/app9152969

73. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Adv Neural Inf Process Syst*. 2021;34:2136–2147. doi:10.48550/arXiv.2106.00908

74. Hammouda K, Khalifa F, El-Melegy M, et al. A deep learning pipeline for grade groups classification using digitized prostate biopsy specimens. *Sensors*. 2021;21(20):6708. doi:10.3390/s21206708

75. Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*. 2021;444:92–110. doi:10.1016/j.neucom.2020.04.157

76. Chen CM, Huang YS, Fang PW, Liang CW, Chang RF. A computer-aided diagnosis system for differentiation and delineation of malignant regions on whole-slide prostate histopathology image using spatial statistics and multidimensional DenseNet. *Med Phys*. 2020;47(3):1021–1033. doi:10.1002/mp.13964

77. Singh M, Kalaw EM, Jie W, et al. Cribriform pattern detection in prostate histopathological images using deep learning models [published online October 9, 2019]. *arXiv*. doi:10.48550/arXiv.1910.04030

78. Silva-Rodríguez J, Colomer A, Sales MA, Molina R, Naranjo V. Going deeper through the Gleason scoring scale: an automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Comput Methods Programs Biomed*. 2020;195:105637. doi:10.1016/j.cmpb.2020.105637

79. Silva-Rodríguez J, Colomer A, Dolz J, Naranjo V. Self-learning for weakly supervised Gleason grading of local patterns. *IEEE J Biomed Health Inform*. 2021;25(8):3094–3104. doi:10.1109/JBHI.2021.3061457

80. Han W, Johnson C, Warner A, et al. Automatic cancer detection on digital histopathology images of mid-gland radical prostatectomy specimens. *J Med Imaging*. 2020;7(4):047501. doi:10.1117/1.JMI.7.4.047501

81. Li J, Li W, Gertych A, Knudsen BS, Speier W, Arnold CW. An attention-based multi-resolution model for prostate whole slide image classification and localization [published online May 30, 2019]. *arXiv*. doi:10.48550/arXiv.1905.13208

82. Eminaga O, Tolkach Y, Kunder C, et al. Deep learning for prostate pathology [published online October 11, 2019]. *arXiv*. doi:10.1038/s42256-020-0200-7

83. Safarpoor A, Hipp JD, Tizhoosh HR. Learning to predict RNA sequence expressions from whole slide images with applications for search and classification [published online March 26, 2022]. *arXiv*. doi:10.48550/arXiv.2203.13997

84. 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. doi:10.1007/s00428-019-02577-x

85. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686–696. doi:10.1038/s41416-020-01122-x

86. Lancellotti C, Cancian P, Savevski V, et al. Artificial intelligence & tissue biomarkers: advantages, risks and perspectives for pathology. *Cells*. 2021;10(4):787. doi:10.3390/cells10040787

87. Toledo-Cortés S, Useche DH, Müller H, González FA. Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression. *Comput Biol Med*. 2022;145:105472. doi:10.1016/j.compbiomed.2022.105472

88. Le Vuong TT, Kim K, Song B, Kwak JT. Joint categorical and ordinal learning for cancer grading in pathology images. *Med Image Anal*. 2021;73:102206. doi:10.1016/j.media.2021.102206

89. Bhattacharjee S, Prakash D, Kim C-H, Choi H-K. Multichannel convolution neural network classification for the detection of histological pattern in prostate biopsy images. *J Korea Multimed Soc*. 2020;23(12):1486–1495. doi:10.9717/kmms.2020.23.12.1486

90. Li Y, Huang M, Zhang Y, et al. Automated Gleason grading and Gleason pattern region segmentation based on deep learning for pathological images of prostate cancer. *Proc 2019 41st Ann Intl Conf IEEE Engr Med Bio Soc*. 2020;8:117714–117725. doi:10.1109/ACCESS.2020.3005180

91. Poojitha UP, Sharma SL. Hybrid unified deep learning network for highly precise Gleason grading of prostate cancer. *Proc 2019 41st Ann Intl Conf IEEE Engr Med Bio Soc*. 2019:899–903. doi:10.1109/EMBC.2019.8856912

92. Bhattacharjee S, Ikromjanov K, Hwang Y-B, Sumon RI, Kim H-C, Choi H-K. Detection and classification of prostate cancer using dual-channel parallel convolution neural network. *Proc Future Tech Conf*. 2021;2:66–83. doi:10.1007/978-3-030-89880-9_6

93. Salman ME, Çakar GÇ, Azimjonov J, Kösem M, Cedimoğlu İH. Automated prostate cancer grading and diagnosis system using deep learning-based Yolo object detection algorithm. *Expert Syst Appl*. 2022;201:117148. doi:10.1016/j.eswa.2022.117148

94. Dudin O, Mintser O, Sulaieva O. Artificial intelligence and next generation pathology: towards personalized medicine. *Proc Shevchenko Sci Soc Med Sci*. 2021;65(2):68–87. doi:10.25040/ntsh2021.02.07

95. Krajňanský V, Gallo M, Nenutil R, Němeček M, Holub P, Brázdil T. Shedding light on the black box of a neural network used to detect prostate cancer in whole slide images by occlusion-based explainability. *bioRxiv*. Preprint posted online April 1, 2022. doi:10.1101/2022.03.31.486599

96. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeogr*. 2008;17(2):145–151. doi:10.1111/j.1466-8238.2007.00358.x

97. Bulten W, Balkenhol M, Belinga J-JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol*. 2021;34(3):660–671. doi:10.1038/s41379-020-0640-y

98. Egevad L, Swanberg D, Delahunt B, et al. Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. *Virchows Arch*. 2020;477(6):777–786. doi:10.1007/s00428-020-02858-w

99. 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. doi:10.1038/s41379-020-0551-y

100. Plazas M, Ramos-Pollán R, León F, Martínez F. Towards reduction of expert bias on Gleason score classification via a semi-supervised deep learning strategy. *Proc SPIE Med Imaging Digit Pathol*. 2022;12032:710–717. doi:10.1117/12.2611517

101. 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;50:167–180. doi:10.1117/12.2611517

102. 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(1):48. doi:10.1038/s41746-019-0112-2

103. Bulten W, Kartasalo K, Chen PHC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154–163. doi:10.1038/s41591-021-01620-2

104. George RS, Htoo A, Cheng M, et al. Artificial intelligence in prostate cancer: definitions, current research, and future directions. *Urol Oncol*. 2022;40(6):262–270. doi:10.1016/j.urolonc.2022.03.003

105. Huang W, Jain P, Randhawa R, et al. AI powered platform to identify primary prostate cancer patients with high risk of recurrence. *Cancer Res*. 2020;80(16)(suppl):2097. doi:10.1158/1538-7445.AM2020-2097

106. Ba W, Wang S, Shang M, et al. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. *Mod Pathol*. 2022;35(9):1262–1268. doi:10.1038/s41379-022-01073-z

107. Ryu HS, Jin M-S, Park JH, et al. Automated Gleason scoring and tumor quantification in prostate core needle biopsy images using deep neural networks and its comparison with pathologist-based assessment. *Cancers*. 2019;11(12):1860. doi:10.3390/cancers11121860

108. Tsuneki M, Abe M, Kanavati F. A deep learning model for prostate adenocarcinoma classification in needle biopsy whole-slide images using transfer learning. *Diagnostics*. 2022;12(3):768. doi:10.3390/diagnostics12030768

109. Ström 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. doi:10.1016/S1470-2045(19)30738-7

110. Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digital Health*. 2020;2(8):e407–e416. doi:10.1016/S2589-7500(20)30159-X

111. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*. 2019;56(4):24–31. doi:10.1109/MSPEC.2019.8678513

112. Bashashati A, Goldenberg SL. AI for prostate cancer diagnosis—hype or today's reality? *Nat Rev Urol*. 2022;19(5):261–262. doi:10.1038/s41585-022-00583-4

113. Hassan T, Hassan B, ElBaz A, Werghi N. A dilated residual hierarchically fashioned segmentation framework for extracting Gleason tissues and grading prostate cancer from whole slide images. *2021 IEEE Sensors App Symp*. 2021:1–6. doi:10.1109/SAS51076.2021.9530155

114. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal*. 2021;67:101813. doi:10.1016/j.media.2020.101813

115. da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol*. 2021;254(2):147–158. doi:10.1002/path.5662

116. Perincheri S, Levi AW, Celli R, et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod Pathol*. 2021;34(8):1588–1595. doi:10.1038/s41379-021-00794-x

117. Jung M, Jin M-S, Kim C, et al. Artificial intelligence system shows performance at the level of uropathologists for the detection and grading of prostate cancer in core needle biopsy: an independent external validation study. *Mod Pathol*. 2022;35(10):1449–1457. doi:10.1038/s41379-022-01077-9

118. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775–784. doi:10.1038/s41591-021-01343-4

119. Aljuhani A, Srivastava A, Cronin JP, Chan J, Machiraju R, Parwani AV. Whole slide imaging: deep learning and artificial intelligence. In: Parwani AV, ed. *Whole Slide Imaging: Current Applications and Future Directions*. Springer; 2022:223–236. doi:10.1007/978-3-030-83332-9_13

120. Dzaparidze G, Kazachonok D, Laht K, Taelma H, Minajeva A. Pathadin—the essential set of tools to start with whole slide analysis. *Acta Histochem*. 2020;122(7):151619. doi:10.1016/j.acthis.2020.151619