Towards a Clinically Useful AI Tool for Prostate Cancer Detection: Recommendations from a PANDA Dataset Analysis

T. J. Hart, Chloe Engler Hart, Aaryn S. Frewing, Dr. Paul M. Urie M.D./Ph.D., Dr. Dennis Della Corte Dr. rer. nat.*

Department of Physics and Astronomy, Brigham Young University, Provo.

Abstract

Objectives: Evaluate the gland-level annotations in the PANDA Dataset and provide specific recommendations for the development of an improved prostate adenocarcinoma dataset. Provide insight into why currently developed artificial intelligence (AI) algorithms designed for automatic prostate adenocarcinoma detection have failed to be clinically applicable.

Methods: A neural network model was trained on 5009 Whole Slide Images (WSIs) from PANDA. One expert pathologist repeatedly performed gland-level annotations on 50 PANDA WSIs to create a test set and estimate an intra-pathologist variability value. Dataset labels, expert annotations, and model predictions were compared and analyzed.

Results: We found an intra-pathologist accuracy of 0.83 and Prevalence-Adjusted Bias-Adjusted Kappa (kappa) value of 0.65. The model predictions and dataset labels comparison yielded 0.82 accuracy and 0.64 kappa. The model predictions and dataset labels showed low concordance with the expert pathologist.

Conclusion: Simple AI models trained on PANDA achieve accuracies comparable to intra-pathologist accuracies. Due to variability within or between pathologists these models will unlikely find clinically application. A shift in dataset curation must take place. We urge for the creation of a dataset with multiple annotations from a group of experts. This will enable AI models, trained on this dataset, to produce panel opinions which augment pathological decision making.

Key Points
1. Current publicly available prostate adenocarcinoma datasets are insufficient to train a clinically useful AI tool capable of accurate Gleason pattern annotations.
2. Collaboration with pathologists could generate a high-quality prostate adenocarcinoma dataset for public use.
3. Training AI algorithms on a distribution of Gleason pattern values instead of single “ground truth” label is a promising avenue towards clinical generalizability.

Introduction

Advances in image recognition enable AI algorithms to better classify and annotate images in pathology. This has spurred the development of algorithms to assist pathologists in detecting and diagnosing prostate adenocarcinoma. Future AI assistants could augment the basic task of image annotation and Gleason pattern assignment.

Prostate adenocarcinoma due to its relatively consistent morphology and associated tissue classification system. Training AI algorithms to detect adenocarcinoma in Hematoxylin & Eosin-stained Whole Slide Images (WSIs)
is simplified by the Gleason grading. To date, multiple AI algorithms have been reported with nearly perfect accuracies of adenocarcinoma. However, only a few prostate adenocarcinoma detecting algorithms are being which questions the generalizability of the developed algorithms.

We hypothesize that the current progress of algorithm development is hindered by the lack of publicly available WSIs with expert gland-level Gleason pattern annotations. Currently, only a small number of publicly available datasets with image-level Gleason grade assignments exist. The largest publicly available dataset of prostate adenocarcinoma WSIs is the PANDA dataset. The PANDA dataset contains 10,616 WSIs each with an assigned image-level Gleason grade. The distribution of Gleason grades in the dataset is roughly equivalent for each category (grades 1-5).

The PANDA dataset lacks gland-level annotations made by expert pathologists — they were done by an algorithm which raises serious concerns about the accuracy of such annotations. Additionally, we found roughly 25% of the annotations are less than 20 pixels in size and are therefore not of clinical relevance. Another concern is the generalizability of the training data which only comes from two different institutions, one institute assigned ~5000 images an overall Gleason grade by one pathologist while the other gave Gleason grades based on pathology reports.

Other datasets include: the TCGA (The Cancer Genome Atlas) dataset, which contains 1,172 WSIs, without any gland-level annotations; the Peso Dataset which has 99 WSIs available with gland level annotations; the SICAPv2 Dataset with 182 WSIs accompanied by annotations of global Gleason scores and path-level Gleason grades; the Harvard Dataset, comprising 71 WSIs.

Here, we report the accuracies of a neural network trained on gland-level annotations from the PANDA dataset with expert annotations. Our findings will motivate the call for a better set of WSI annotations and shift in strategy for training next generation AI models for clinical use.

**Methods**

**Data Collection and Preparation**

5065 WSIs from PANDA dataset (Radboud-produced) were divided into training (90%), validation (9%), and test sets (1%). The images were tiled within each subset into 512 x 512 tiles, padding WSIs where necessary.

**Model Training and Evaluation**

A ResNet-50 model (PointRend) was trained within the Detectron2 framework. The hyperparameters searched included learning rate, batch size, and image dimension. Additionally, various sampling strategies were implemented, including oversampling of tiles containing benign, GP 3, GP 4, and GP 5 labels and augmentations such as rotation and mirroring of tiles. The validation set was used as the termination criterion for training and for model comparison. The model which used learning rate schedule from 0.001 to 0.0001, batch size 16, image dimension 512, and oversampling to balance classes benign, GP 3, GP 4, and GP 5 exhibited the best validation score and was selected for evaluation on the test set.

**Pathologist Annotations**

Annotations followed a two-step process involving both a pre-med student and an experienced pathologist. First, the pre-med student was trained in tumor identification. He then annotated regions in each WSI that were deemed cancerous. Subsequently, the pathologist reviewed the student’s annotations, corrected any inaccuracies, and added new regions as necessary to ensure comprehensive coverage of the Gleason patterns.

Later, the student replicated the initial annotations while intentionally introducing modifications to make them unfamiliar to the pathologist. These modifications included erasing all annotations in certain WSIs or assigning an incorrect classification to all tumor regions within a WSI. The student then presented these modified WSIs to the pathologist — who was unaware that they were the same WSIs previously annotated — for re-annotation.

**Annotation Comparison with Pathologists and Statistical Evaluation**

The annotations made by the model, test set, and pathologists, were compared on regions of interest (ROI) labeled as benign, GP 3, GP 4, or GP 5 by the initial annotations of the pathologist.

Using the specified regions, confusion matrices were generated from pixel-wise comparisons for each pair combination of the annotation masks, resulting in six confusion matrices (e.g. pathologist round 1 vs model, etc.). We calculated the accuracy and Prevalence-Adjusted Bias-Adjusted Kappa (kappa21) values for each WSI annotation. After the accuracy and kappa values were calculated for each annotation mask of each WSI, all the accuracies and kappa for each WSI comparison combination were averaged (e.g. an average accuracy and average kappa was found for pathologist round 1 vs model). Thus, six mean accuracies and six mean kappa values were obtained.

Citation: *Dennis Della Corte. Towards a Clinically Useful AI Tool for Prostate Cancer Detection: Recommendations from a PANDA Dataset Analysis*. Jour of Clin Cas Rep, Med Imag and Heal Sci 5(3)-2023.

DOI: 10.55920/JCRMHS.2023.05.001216
Figure 1: Binary comparison of regions of interest between pathologist annotation rounds (Patho 1 and Patho 2), annotation labels from PANDA test set (Labels), annotation from network inferences (Model). Agreement in green, disagreement in red. Kappa (K) - see methods for details – and accuracy (Acc) for all six combinations are shown below each panel.

Key Message:
Models can never get better than labels, and due to the inter-pathologist difference labels have low accuracy. Even a simple model can learn the labels. Better datasets are needed, not better AI models.
**Figure 2:** Provides an overview of the features expected in the ideal prostate adenocarcinoma dataset. Each feature is highlighted in the discussion section.
Results

The model yields an accuracy of 0.82 and kappa of 0.64 when evaluated on the test set labels from PANDA. Similar values are found between pathologist’s two sets of annotations (as depicted in Fig. 1). It is worth noting that the comparison between the pathologist’s first and second annotations was conducted using the same specified regions employed in the model’s evaluation on the test set. A low level of agreement between the model or labels and the pathologist annotations (all the combinations of comparisons provided similar accuracy/kappa values, refer to Fig. 1) can be seen.

Discussion

Existing gland-level annotations in the PANDA dataset are very different from the annotations made by an expert pathologist. As such, they cannot be reliably used to train AI algorithms – unsurprisingly, given that the Radboud data was annotated by a different AI algorithm. Interestingly, when the Pathologist performed visual quality control on the Radboud labels, no major issues were apparent. The inferiority of the dataset was only highlighted by our systematic test which achieved worse than random kappa values.

The biggest issues in generating gland-level annotations are inter- and intra-pathologist variation.22 In absence of specific mathematically grounded criteria, the Gleason grading system allows a degree of subjectivity on the part of the pathologist in each grade assignment. This variation frustrates the ability to label images with certainty, which carries over into the performance of any AI algorithm.

To reconcile different gland-level annotations, individual pathologist labels have been averaged and merged into one single ground truth.23 Unfortunately, this method of labeling does not replicate the variation seen in clinical practice between expert pathologists. Therefore, instead of establishing one single ground truth it seems more appropriate to create a distribution of possible Gleason Patterns for each individual gland, and then use those distributions to train algorithms.

The current assumption in AI development for prostate adenocarcinoma detection is that AI should replace a super-expert pathologist who always gets assignments right. However, this assumption is flawed as pathologists do not agree on gland-level annotations.24 We propose an alternative approach: training AI models to predict the distribution of responses from a panel of expert pathologists. These predictions would assign a distribution over all possible Gleason pattern scores. Clinical pathologists could then use these predicted annotations to gain insights like consulting multiple colleagues. Training such a model would require costly and laborious annotation of each image by different pathologists at the gland level.

Based on our findings and previous research7 we put forth criteria for what would constitute the ideal high-quality prostate adenocarcinoma pathology dataset, as depicted in Figure 2. The dataset should be comprised of full-size WSIs, because they capture the entire tissue section at high resolution and allow analysis of tissue structures, cell morphology, and other relevant features, as opposed to patches or pixel clusters. The ideal dataset should be sufficiently sized for an algorithm to train on. We estimate this significant number to be 20,000 WSIs. Sufficient variation is necessary in the following three categories: patient demographics, prostate adenocarcinoma type; and adenocarcinoma severity.25 Finally, the dataset should be easily accessible to the public. It should be organized and stored in a consistent vendor agnostic format that allows researchers to retrieve and use the data efficiently. Providing open access or appropriate permissions for accessing the dataset encourages collaboration, accelerates research progress, and enables the development of innovative techniques for prostate adenocarcinoma diagnosis and treatment.

Conclusion

It is evident that addressing the limitations associated with current datasets is crucial in advancing the field of AI in prostate adenocarcinoma pathology. As highlighted, the reliance on proprietary datasets poses several challenges, including limited access, potential bias, and reduced generalizability. To overcome these obstacles, collaboration on a public dataset emerges as a promising solution. By pooling together diverse and comprehensive datasets from multiple sources, a public dataset would foster improved accuracy and reliability of machine learning models. Furthermore, a public dataset would enhance transparency, promote standardized evaluation methods, and facilitate the reproducibility of research findings. We also suggest that the application of panel annotations without combining them into a ground truth label will yield more clinically relevant AI algorithms.

Acknowledgement of Sources of Support: We would like to thank Brigham Young University for their ongoing financial support for the undergraduate students involved in the research project.

Author Acknowledgements: TJH and CEH were primarily responsible for the data analysis and model evaluation. PU and AF developed the dataset of test images. DDC conceived the study, secured funding, and mentored students. All authors supported the writing process.
References


