Modern Pathology (2021) 34:478–489
https://doi.org/10.1038/s41379-020-00674-w

ARTICLE

High throughput assessment of biomarkers in tissue microarrays using artificial intelligence: PTEN loss as a proof-of-principle in multi-center prostate cancer cohorts

Stephanie A. Harmon[1,2] · Palak G. Patel[3,4,5] · Thomas H. Sanford[1,6] · Isabelle Caven[3,4] · Rachael Iseman[3,4] · Thiago Vidotto[7] · Clarissa Picanço[7] · Jeremy A. Squire[4,7] · Samira Masoudi[1] · Sherif Mehralivand[1] · Peter L. Choyke[1] · David M. Berman[3,4] · Baris Turkbey[1] · Tamara Jamaspishvili[3,4]

Received: 30 December 2019 / Revised: 21 August 2020 / Accepted: 21 August 2020 / Published online: 3 September 2020

Abstract

Phosphatase and tensin homolog (PTEN) loss is associated with adverse outcomes in prostate cancer and has clinical potential as a prognostic biomarker. The objective of this work was to develop an artificial intelligence (AI) system for automated detection and localization of PTEN loss on immunohistochemically (IHC) stained sections. PTEN loss was assessed using IHC in two prostate tissue microarrays (TMA) (internal cohort, $n = 272$ and external cohort, $n = 129$ patients). TMA cores were visually scored for PTEN loss by pathologists and, if present, spatially annotated. Cores from each patient within the internal TMA cohort were split into 90% cross-validation ($N = 2048$) and 10% hold-out testing ($N = 224$) sets. ResNet-101 architecture was used to train core-based classification using a multi-resolution ensemble approach (×5, ×10, and ×20). For spatial annotations, single resolution pixel-based classification was trained from patches extracted at ×20 resolution, interpolated to ×40 resolution, and applied in a sliding-window fashion. A final AI-based prediction model was created from combining multi-resolution and pixel-based models. Performance was evaluated in 428 cores of external cohort. From both cohorts, a total of 2700 cores were studied, with a frequency of PTEN loss of 14.5% in internal (180/1239) and external 13.5% (43/319) cancer cores. The final AI-based prediction of PTEN status demonstrated 98.1% accuracy (95.0% sensitivity, 98.4% specificity; median dice score = 0.811) in internal cohort cross-validation set and 99.1% accuracy (100% sensitivity, 99.0% specificity; median dice score = 0.804) in internal cohort test set. Overall core-based classification in the external cohort was significantly improved in the external cohort (area under the curve = 0.964, 90.6% sensitivity, 95.7% specificity) when further trained (fine-tuned) using 15% of cohort data (19/124 patients). These results demonstrate a robust and fully automated method for detection and localization of PTEN loss in prostate cancer tissue samples. AI-based algorithms have potential to streamline sample assessment in research and clinical laboratories.

Introduction

Modern artificial intelligence (AI) techniques have demonstrated the potential to achieve human-level

These authors contributed equally: Stephanie A. Harmon, Palak G. Patel

Supplementary information The online version of this article (https://doi.org/10.1038/s41379-020-00674-w) contains supplementary material, which is available to authorized users.

✉ Tamara Jamaspishvili
tamara.jamaspishvili@queensu.ca

[1] Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

[2] Clinical Research Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA

[3] Division of Cancer Biology & Genetics, Cancer Research Institute, Queen's University, Kingston, ON, Canada

[4] Department of Pathology and Molecular Medicine, Queen's University, Kingston, ON, Canada

[5] Department of Cell Biology at The Arthur and Sonia Labatt Brain Tumour Research Centre at the Hospital for Sick Children, Toronto, ON, Canada

[6] Department of Urology, Upstate Medical University, Syracuse, NY, USA

[7] Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

performance in various computer vision and medical imaging applications. Specifically applied to digital pathology, deep learning/convolutional neural networks have shown promising accuracy for automated detection and grading of several disease types [1–7]. Recent literature supports the potential for AI-based biomarker assessment, where deep learning algorithms have shown high accuracy in automated human epidermal growth factor receptor 2 scoring in breast cancer and programmed death receptor 1 scoring in lung cancer [8–10]. Well-trained AI applications in IHC offer the potential for reproducible immunoscoring and quantitation to assist biomarker assessment without additional burden to pathologists, where subjective scoring and technical staining quality often lead to variation in human interpretation [11–14].

Phosphatase and tensin homolog (PTEN), a tumor suppressor gene, is a crucial regulator of the oncogenic PI3K/AKT/mTOR signaling pathway and its loss of function is one of the most common events observed in many types of cancer [15, 16]. Genomic aberrations of PTEN or protein loss are among the most common in prostate cancer and have been shown to be associated with aggressive prostate cancer and unfavorable patient outcomes after definitive local therapy [17–20]. Current guidelines for prognostication of localized prostate cancer are driven solely by assessment of clinico-pathological parameters such as age, prostate-specific antigen, cancer grades, and stage [21, 22]. Clinically established risk stratification algorithms alone lack personalized risk assessments especially in the group of men characterized as low and intermediate risk, commonly leading to overtreatment or undertreatment of the disease [23]. To address these gaps and improve risk stratification and treatment management of prostate cancer patients, prognostic information from the molecular biomarkers or genomic classifiers should be integrated into the standard clinical parameters if that would impact on a short- or long-term clinical management [24, 25]. Numerous studies have reported that the use of PTEN loss as a prognostic biomarker can provide clinically relevant information at a lower cost since the development of efficient clinical-grade immunohistochemistry (IHC) assay, making it an attractive biomarker of aggressive disease in a clinical workflow [17, 20, 26, 27].

PTEN loss is a well-studied biomarker in prostate cancer which might be soon integrated into the clinical practice. As a proof-of-principle, here we use digital images of PTEN IHC to develop and validate a deep learning-based workflow for automated detection and spatial annotation of PTEN loss in tissue microarrays (TMA). In addition, we assess the generalization of this approach by performing validation using tumors on a TMA from an external patient population.

Methods

Patients and cohorts

We used prostate TMA from two institutions: Kingston Health Services Center (KHSC), Canada ($n = 272$, RP years, 2000–2012), as an internal cohort, and the University of Sao-Paulo, Brazil ($n = 129$; RP years, 2006–2015), as an external cohort. High-density TMAs from archival surgical tissues contained five 0.6 mm cancer cores/per case and four benign cores/per case on average for KHSC TMAs and three 1.0 mm cancer cores/per case and one benign core/per case for the Brazilian TMAs. Clinical and pathological information are provided in Supplementary Table S1 and are further detailed in previous publication [28].

Immunohistochemical staining (IHC), slide scoring, and pathologists' manual annotations

For both KHSC and Brazilian cohorts, IHC staining was performed on an automated staining platform Discovery XT (Ventana Medical System, Inc., Tucson, AZ, USA). Briefly, TMA blocks were sectioned at 5 µm and stained with rabbit monoclonal anti-PTEN antibody (Clone-D4.3 XP, dilution-1:100, Cell Signaling Technologies). Staining conditions are further detailed in Supplementary Table S2.

Stained TMA were scanned at ×20 on an Aperio scanner (Leica Biosystems). Protein expression was independently scored by two urologic pathologists (TJ and DMB) using proposed scoring criteria [20, 29–31]. Stromal cells and benign glands were utilized as internal positive controls. In cancer cells, intact PTEN was defined as cytoplasmic and/or nuclear staining above background. PTEN loss was defined as complete (100% of sampled tumor cells) or partial (<100%) loss of cytoplasmic and/or nuclear staining. Examples of complete PTEN loss and partial PTEN loss are shown in Fig. 1. "Low PTEN" was defined as cancer cells showing significantly diminished PTEN protein expression compared to an internal positive control (either benign epithelium or stroma) (Supplementary Fig. S1). Any cores with "low" PTEN expression still remained in the study if appropriate positive control expression was identified in either benign or stromal regions in the given case. Any cores with substantial tissue, staining or scanning artifacts noted by the pathologists were excluded from analysis, resulting in a total patient population of $n = 271$ for the internal (KHSC) cohort and $n = 124$ for the external (Brazilian) cohorts. Tumor regions with PTEN loss (i.e., regions of interests (ROIs)) were identified visually and manually annotated by a pathologist. Annotations used for training were performed by pathologists within PTEN loss containing cancer cores.

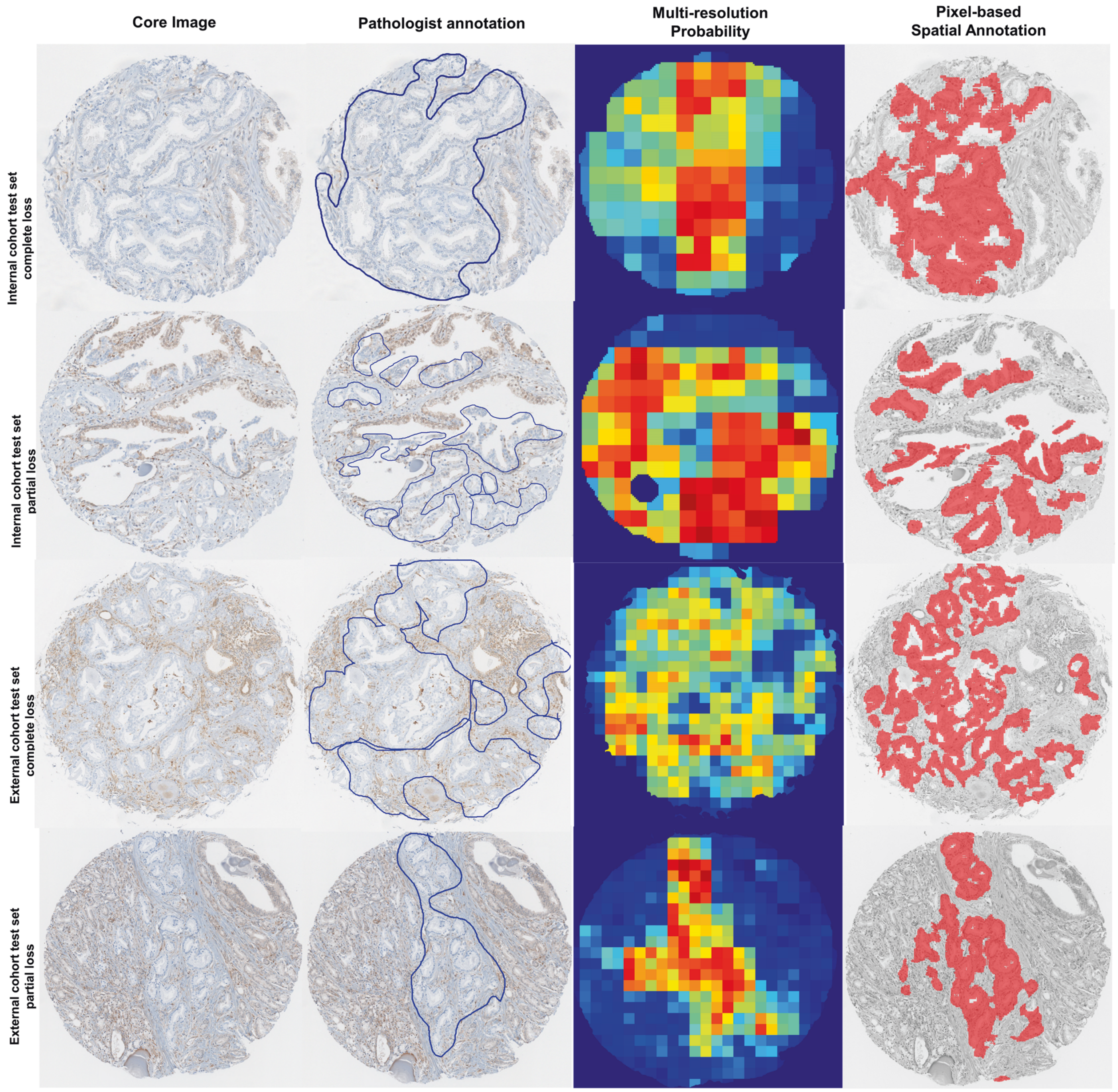
Core Image
Pathologist annotation
Multi-resolution Probability
Pixel-based Spatial Annotation
Internal cohort test set complete loss
Internal cohort test set partial loss
External cohort test set complete loss
External cohort test set partial loss

Fig. 1 Example true positive (correctly identified as having PTEN loss) cases from internal and external testing cohorts. Note external probability maps show performance of the algorithm after fine-tuning. Top: internal cohort testing data set; case with complete PTEN loss in tumor cells, multi-resolution probability = 43.4% and dice = 0.738. Middle top: internal cohort testing data set; case with partial PTEN loss, multi-resolution probability = 56.6% and dice = 0.761. Middle bottom: external cohort testing data set; case with complete PTEN loss in tumor cells, multi-resolution probability = 37.7% and dice = 0.552. Bottom: external cohort testing data set; case with partial PTEN loss in tumor cells, multi-resolution probability = 27.7% and dice = 0.347. Cores with probability > 26.5% were classified as "PTEN loss" in multi-resolution approach. Cores with any pixel region classified as PTEN loss in binary mask were classified as "PTEN loss" in pixel-based approach.

Data sets for deep learning algorithm development and evaluation

Two different approaches were used in this study: (1) a multi-resolution approach for automated identification of TMA cores containing PTEN loss and (2) a pixel-based approach for automated spatial localization of the regions with PTEN loss within TMA cores. The entire image processing and classification pipeline is shown in Fig. 2 and training assignments for each patient cohort are summarized in Table 1.

We trained and evaluated our algorithms in three steps: (1) all cores ($n = 2272$ cores, $N = 271$ patients) from internal cohort TMAs (KHSC) were randomly divided into training and testing sets. This assignment was done on the patient level, meaning all cores belonging to an individual

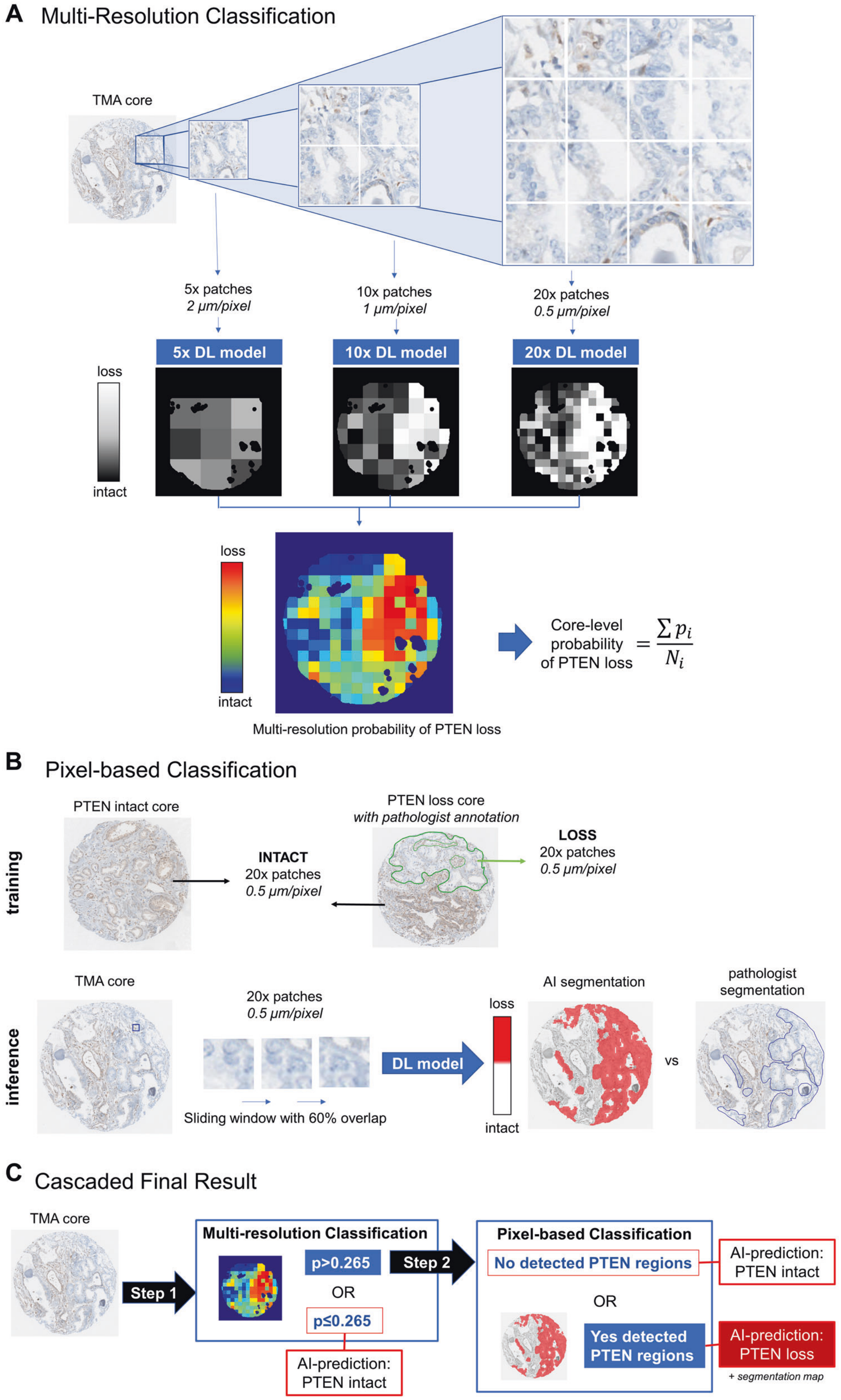

Fig. 2 Deep learning workflow for characterization of TMA cores. a Core-base classification using multi-resolution data trained from core-level label (no spatial annotation). The result of the multi-resolution approach is a core-level probability of PTEN loss derived by pixel-based averages of ×5, ×10, and ×20 algorithm predictions. **b** Pixel-based classification at ×20 trained from pathologist spatial annotation of PTEN loss regions. The result of the pixel-based classification is a spatial map of pixel predicted as containing PTEN loss cells, derived from average of sliding window-based inference. **c** The final model consisted of a combined (cascaded) approach of (**a**) followed by (**b**). If a core-level probability of PTEN loss from the multi-resolution based approach was above a determined threshold (0.265), the core would be sent to the pixel-based classification approach. Only cores containing areas predicted to have PTEN loss by both algorithms (above threshold and included spatial area of PTEN loss) received a final AI-based prediction of PTEN loss.

patient were assigned to either training or testing set with no overlap. The internal training set included $n = 2048$ cores from $N = 243$ patients. Stratified cross-validation was used to iteratively split this training set into five folds on the patient level (Table S3). The internal hold-out testing set was composed of $n = 224$ cores from $N = 243$ patients

Table 1 Core-based distribution of training, validation, and testing sets for internal cohort and external cohort.

Data set	Subset	Patients (*N*)	Cores (*N*)	Cancer (%)	PTEN loss (%)
Internal cohort	Cross-validation	243	2048	1099 (54%)	161 (15%)
	Hold-out testing set	28	224	134 (60%)	19 (13%)
External cohort	–	124	428	319 (75%)	43 (13%)
External cohort fine-tune	Training set	14	52	38 (73%)	9 (23%)
	Validation set	5	19	14 (74%)	2 (14%)
	Hold-out testing set	105	357	267 (75%)	32 (12%)

Cores were randomly split into sets at the patient level, i.e., all cores from individual patient were stratified to each subset. Proportion of cores containing cancer is listed, as well as number of cancer cores with PTEN loss. All cores, both cancer and benign, were utilized for training and evaluation. For fine-tuning in external cohort, the model trained from internal data was retrained using $n = 71$ cores from external cohort and tested on remaining 357 cores. Hold-out testing refers to data that are not included in any training or validation procedures.

(note: hold-out testing refers to data that are not included in any training or validation procedures). Results from the internal validation and internal test sets were reported separately, (2) all external TMA core data (Brazilian cohort) were used as a separate independent testing set ($n = 428$ cores, $N = 124$ patients), (3) algorithms from cross-validation training cohort were fine-tuned using 15% of the external TMA data, with random selection of $N = 19$ patients (71 cores). Then, the fine-tuned algorithms were applied to the remaining 85% of external TMA data ($n = 357$ cores, $N = 105$ patients).

Both benign and cancer cores were included in training and testing analysis. In all steps, all core images from an individual patient were not split between sets (i.e., training, validation, or testing sets). During validation, no pathologist's annotations were required for input.

Multi-resolution approach for core-based classification

Each core image was extracted at ×20 from the TMAs resulting in 2000 × 2000 pixels/core image for internal cohort (0.6 mm diameter cores) and 2800 × 2800 pixels/core image for external cohort (1 mm diameter cores). Patches of 100 × 100 pixels were extracted from ×5, ×10, and ×20 objectives and were included for image processing. Deep learning models (ResNet-101 architecture) were trained from patches at ×5, ×10, and ×20 objectives, respectively (Fig. 2a). All patches were labeled according to the PTEN status of the core. The final training parameters are included in Supplementary Table S4. All models were trained using fastai library (https://github.com/fastai/fastai). After completion of training, models were applied to validation and testing sets and a multi-resolution map of core image was generated, where every pixel represented average probability of PTEN loss from each resolution model (Fig. 2a). The average probability of PTEN loss from all pixels within an entire core image was reported as the AI-based score.

Due to the well-known differences in patient populations across different medical centers, as well as differences in staining and tissue processing in clinical and research laboratories, it is possible that a model trained at one institution may not achieve high performance when directly applied to a new patient cohort. Therefore, an additional training method, fine-tuning was used to enhance performance of models using 15% of the patients from external cohort (Table 1). For this process, pretrained cross-validation models from internal cohort data were used to initialize weights for fine-tuning on external training data (Supplementary Table S4). Following training, models were deployed to all remaining cores of external testing set ($N = 357$).

Pixel-based approach for spatial annotation

All image patches were derived in reference to 50 × 50 pixel regions at ×20 objective within pathologically annotated regions of PTEN loss. Patch locations were determined by fitting the minimum number of nonoverlapping regions that fully contained pathologist annotations (Fig. 2b). All remaining patches produced beyond pathologist annotations were labeled as "PTEN intact." A ResNet-101 architecture was trained from patches resampled to 100 × 100 pixels (simulated ×40 resolution). The model was then applied using a sliding-window approach with 60% overlap between neighboring patches (Fig. 2b). The average probability of PTEN loss was generated at each pixel location in the core image. A refined binary mask was created from a threshold-based mask (>50% probability of PTEN loss) with additional post-processing for identification of distinct morphological areas, with excluded regions defined as those <0.00125 mm^2 or those with maximum probability of PTEN loss < 0.75. The final binary prediction mask assigns pixels with PTEN loss = 1 and all other regions = 0. Full details are provided in Supplementary Material. All post-processing and image analysis were performed in MATLAB (R2018b, https://www.mathworks.com).

Statistical analysis for deep learning algorithm evaluation

For the multi-resolution approach, the AI-based probability of PTEN loss per core was evaluated and area under the curve (AUC) of ROC (receiver-operating curve) analysis was reported. Cross-validation performance was reported for each fold. A probability cut-off, defined as threshold maximizing specificity while achieving 95% sensitivity (at least 95%), was determined from the internal cross-validation performance and set at 0.265. Accuracy, sensitivity, and specificity at the identified threshold were reported for correct classification of PTEN status for each core. In testing sets, the reported AI-based probability used for performance metric calculation was the average AI probability from all cross-validation models. An individual core was labeled as "PTEN loss" if the multi-resolution probability was >26.5%. For performance metrics, each core was defined as one of the following: true positive refers to correct prediction of PTEN loss, true negative refers to correct prediction of PTEN intact, false positive refers to incorrect prediction of PTEN loss, and false negative referred to incorrect prediction of PTEN intact.

For the pixel-based approach, the Sorensen–Dice coefficient was used to calculate the pixel-based similarity of AI-based binary mask vs. pathologist spatial annotation. Dice is defined as twice the area of overlap between regions divided by sum of total area of both regions (see Supplementary Material). In testing sets, the predicted probability for PTEN loss in each spatial patch within a core was averaged across all cross-validation models. An individual core was labeled as "PTEN loss" if the binary detection mask included any pixel regions with value = 1. Accuracy, sensitivity, and specificity of cores with PTEN loss detection were reported.

A final AI-based prediction was created from combining (cascading) multi-resolution and pixel-based models, where an individual core was considered as containing PTEN loss if the multi-resolution average probability was above the predetermined threshold (0.265) and pixel-based binary detection mask contained regions labeled as PTEN loss (i.e., the core was labeled as "PTEN loss" by both algorithms).

Definitions of all performance metrics are provided in Supplementary Material. Ninety-five percent confidence intervals and standard errors of the prediction performance metrics were calculated from 2000 bootstrap samples by randomly sampling patients with replacement. AI-based quantitative metrics (multi-resolution average probability, pixel-based dice coefficient) were evaluated across qualitative levels of PTEN loss (i.e., intact vs. intact low and partial loss vs. complete loss) using Wilcoxon rank-sum test using the Rosner–Glynn–Lee method to account for multiple cores per patient. All statistical analysis was performed in R (version 3.4.1).

Results

In total, 2272 prostate cores from prostate cancer patients were included in the internal cohort, split into 90% cross-validation training and 10% testing ($N = 224$) sets (Table 1). By pathologist scoring, the overall frequency of PTEN loss for cancer cores was 14.5% (180/1233), where 26.1% (47/180) showed partial PTEN loss and 73.9% (132/180) exhibiting complete PTEN loss. In the external cohort, frequency of PTEN loss was similar at 13.5% (43/319) of which 12/43 had partial PTEN loss and 72.1% (31/43) had complete PTEN loss.

Multi-resolution classification performance

Overall classification performance (AUC) of the multi-resolution approach for the internal cohort was 0.989 (95% CI: 0.980–0.996) and 0.993 (95% CI: 0.975–1.00) in cross-validation and testing sets, respectively. Median cross-validation performance at each resolution ranged from AUC 0.980 to 0.990 (Supplementary Table S4). Since the accuracy is heavily influenced by the large proportion of PTEN intact cores, the probability threshold for determining PTEN loss was optimized based on sensitivity. Using a probability threshold of >26.5% likelihood of PTEN loss to achieve minimum 95% sensitivity in cross-validation, the accuracy was 93.9% (95% CI: 92.2–95.5) and 95.1% (95% CI: 90.9–98.3) in the cross-validation and testing sets, respectively (Table 2). Overall, cross-validation models yielded similar performance, with median 0.991 ± 0.006 standard deviation (Supplementary Table S5). As expected, within cores with PTEN loss, the average probability of PTEN loss was significantly higher in cores with complete loss compared with partial loss in cross-validation ($p = 0.0003$) and, a similar result was also observed in the testing set, though not statistically significant (Fig. 3). Median probability of PTEN loss in cores with complete vs. partial loss was 0.7013 (range: 0.15–0.8796) vs. 0.4578 (range: 0.1783–0.7904) and 0.7446 (range: 0.2670–0.8807) vs. 0.5656 (range: 0.2680–0.7360) in cross-validation and testing sets, respectively. Of eight false negatives cores (incorrectly classified as PTEN intact) in cross-validation, seven had partial PTEN loss and one had focal loss. Review of 11 false positive cores (incorrectly classified as PTEN loss) demonstrated that majority were cancer containing cores (7/11) and of these, 2 cores had low PTEN expression (Supplementary Fig. S1 and Supplementary Table S6) and 7 had relatively low epithelial to stromal ratio, i.e., cores having high stromal content, as assessed qualitatively (Supplementary Fig. S2).

We first kept the probability threshold constant at 26.5%, and applied a multi-resolution based approach to

Table 2 Performance metrics of multi-resolution AI model in internal cohort validation set, testing set, and external cohort testing set.

Cohort	AUC	Average probability threshold $p > 0.265$						
		TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Internal cohort validation set ($N = 2048$)	0.989 (0.980–0.996)	153	1770	117	8	95.03% (90.4–98.4)	93.80% (92.0–95.6)	93.90% (92.2–95.5)
Internal cohort test set ($N = 224$)	0.993 (0.975–1.00)	19	194	11	0	100.00% (100–100)	94.63% (90.3–98.1)	95.09% (90.9–98.3)
External cohort ($N = 428$)	0.963 (0.909–0.994)	42	161	224	1	97.67% (91.1–100)	41.82% (35.9–48.0)	47.43% (41.9–53.3)
External cohort with fine-tuning (test set $N = 357$)	0.964 (0.902–0.998)	29	311	14	3	90.63% (75.0–100)	95.69% (93.6–97.6)	95.24% (93.0–97.3)

External cohort performance is reported for all cores ($N = 428$) and with fine-tuning ($N = 357$). 95% confidence intervals calculated from bootstrap analysis on the patient level. AUC is reported for continuous probability of PTEN loss. TP = correctly classified as PTEN loss; TN = correctly classified as PTEN intact; FP = incorrectly classified as PTEN loss; FN = incorrectly classified as PTEN intact.

the external cohort. Unfortunately, this approach increased the rate of false positive cores, resulting in an AUC of 0.963 and decreased accuracy of 47.34% (95% CI: 41.9–53.3) (Table 2). Specifically, the performance of ×20 cross-validation models decreased the most with mean AUC 0.942 (range: 0.926–0.951), while performance of ×10 cross-validation models remained highest with a mean AUC of 0.972 (range: 0.965–0.974) (Supplementary Table S4). After fine-tuning with 15% of the external cohort data, the overall AUC increased to 0.964 (95% CI: 0.902–0.998) when tested on the remaining cores of the external cohort ($n = 357$) (Table 2). The accuracy increased to 95.2% (95% CI: 93.0–97.3) at the pre-defined 26.5% probability threshold (Table 2). Similar to what was observed for the internal cohort data, the average probability of PTEN loss was significantly higher in cores annotated by pathologists as having complete loss and partial loss, both with ($p = 0.005$) and without ($p = 0.034$) after fine-tuning (Fig. 4). False negativity was due to heterogeneous PTEN staining, e.g., all three cores exhibiting partial PTEN loss. Interestingly, some cores had decreased PTEN staining that did not qualify as complete loss as noted by pathologist. We found that these "PTEN low" cores demonstrated increased average probability of PTEN loss (Fig. 4). With respect to false positive results ($n = 14$ TMA cores), we noted that 5/14 occurred in benign cores exhibiting lower than normal PTEN expression.

Pixel-based classification and spatial annotation performance

In the pixel-based classifier, classification performance was evaluated by detection of any region with AI-predicted PTEN loss in the core. Per core results demonstrated accuracies of 96.5% (95% CI: 95.4–97.6) and 96.4% (95% CI: 92.6–99.1) in the cross-validation and testing sets, respectively (Supplementary Table S7). No false negatives were recorded. Representative examples of AI-based spatial annotation maps within complete and partial PTEN loss cores are shown in Fig. 1. Using the post-processed predictions of PTEN loss regions, dice results were favorable with median 0.811 (range: 0–0.94) and median 0.8043 (range: 0–0.97) in cross-validation and testing sets, respectively (Supplementary Table S7). Evaluating variability in individual cross-validation models demonstrated similar performance after post-processing, with median 0.809 ± 0.0097 standard deviation (Supplementary Table S8).

Applying the pixel-based classifier to all external cores ($n = 428$), accuracy only achieved 66.2% (95% CI: 59.9–72.0) due to high false positive regions in PTEN intact cores (Supplementary Table S7). These false

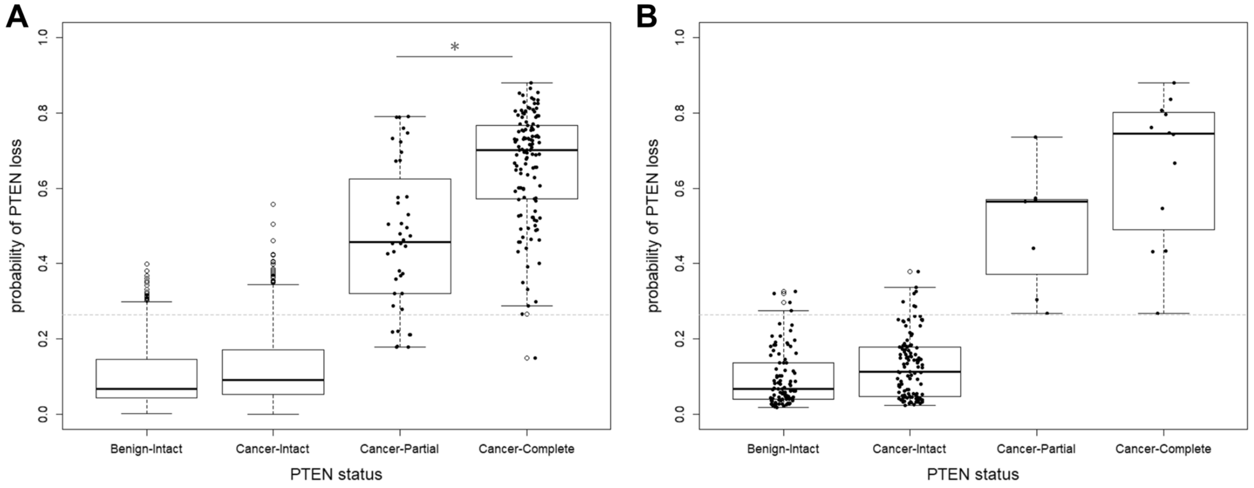

Fig. 3 Core-based average probability of PTEN loss within internal cohort based on pathologist labels. a validation set, $N = 2048$, and **b** testing set, $N = 224$; For training and evaluation purposes, cancer cores with partial and complete loss were grouped as "PTEN loss," while both benign and cancer cores with intact staining were grouped as "PTEN intact." Cores with AI-based probability > 26.5% were classified as "PTEN loss" (dashed gray line). Each individual cores is shown as a datapoint, except in the benign cores of validation set (panel **a**), due to high number of cores in these groups ($n = 949$ benign intact, $n = 938$ benign cancer). Any benign or cancer cores with intact PTEN staining classified as "PTEN loss" represent false positive by AI. Any cancer cores with partial or complete PTEN loss classified as "PTEN intact" represent false negative by AI.

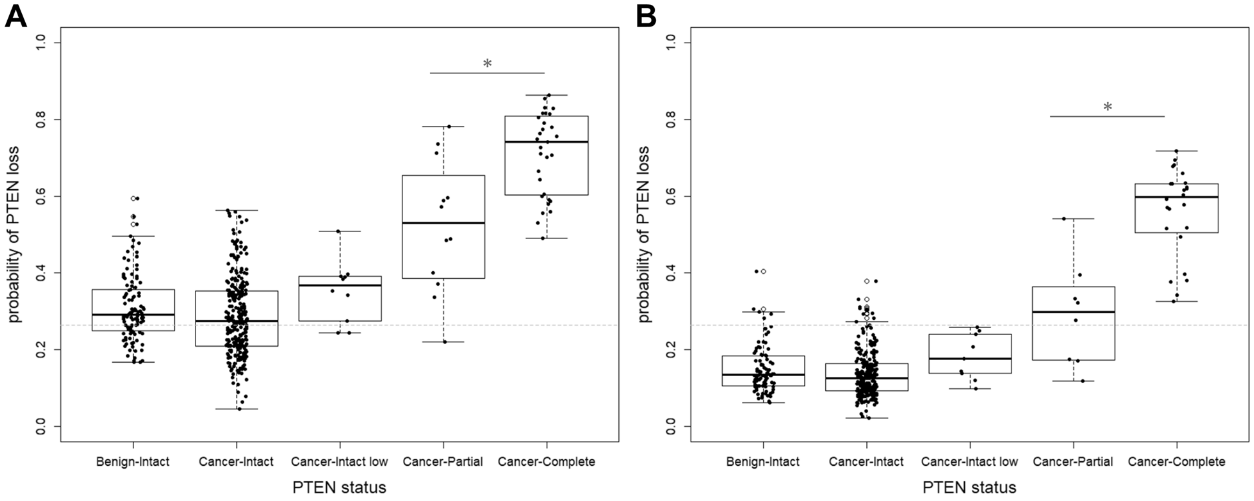

Fig. 4 Core-based average probability of PTEN loss in external cohort with and without fine-tuning. a with no fine-tuning $N = 428$ and **b** after fine-tuning, $N = 357$. For training and evaluation purposes, cancer cores with partial and complete loss cases were grouped as "PTEN loss," while both benign and cancer cores with intact staining were grouped as "PTEN intact." Cancer cores with PTEN intact (low) staining were also included in "PTEN intact" group. Cores with AI-based probability >26.5% were classified as "PTEN loss." Cores with AI-based probability >26.5% were classified as "PTEN loss" (dashed gray line). Each individual cores is shown as a datapoint. Any benign or cancer cores with intact (or low) PTEN staining classified as "PTEN loss" represent false positive by AI. Any cancer cores with partial or complete PTEN loss classified as "PTEN intact" represent false negative by AI.

positives were again observed within cores with lower epithelial/stromal ratio (i.e., higher stromal content) (Supplementary Fig. S2). Again, no false negatives were recorded. Within cores with partial or complete PTEN loss, median dice was 0.7392 (range: 0–0.924) compared to pathologist annotations. Dice was significantly higher in cores with complete PTEN loss, median 0.7779 (range: 0.2305–0.9197), compared with cores with partial PTEN loss, median 0.32034 (0–0.9242), $p = 0.002$.

Combined model performance

The pixel-based approach did not undergo fine-tuning due to the complementary nature of the two models and prior

Table 3 Simulation of high-throughput workflow by sequential combination of multi-resolution and pixel-based algorithms.

Cohort	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Internal cohort validation set ($N = 2048$)	153	1856	31	8	95.03% (90.4–98.4)	98.36% (97.6–99.0)	98.10% (92.6–99.1)
Internal cohort test set ($N = 224$)	19	203	2	0	100.00% (100–100)	99.02% (97.4–100)	99.11% (97.7–100)
External cohort ($N = 428$)	42	263	122	1	97.67% (91.1–100)	68.31% (62.1–73.8)	71.26% (65.6–76.4)
External cohort with fine-tuning (test set $N = 357$)	29	311	14	3	90.63% (75.0–100)	95.69% (93.6–97.6)	95.24% (93.0–97.3)

Any core predicted as PTEN loss by multi-resolution was considered for segmentation by pixel-based algorithm. Only cores predicted as PTEN loss by both algorithms were classified as PTEN loss. External cohort performance is reported for all cores ($N = 428$) and with fine-tuning ($N = 357$). 95% confidence intervals calculated from bootstrap analysis on the patient level. TP = correctly classified as PTEN loss; TN = correctly classified as PTEN intact; FP = incorrectly classified as PTEN loss; FN = incorrectly classified as PTEN intact.

fine-tuning of multi-resolution approach. Therefore, a final combined result was obtained by sequential (cascaded) application of the multi-resolution predictor followed by pixel-based spatial annotations (Table 3). An individual core was classified as having PTEN loss if the multi-resolution average probability was >26.5% and it contained regions suspicious for PTEN loss based on pixel-based classification. The combined method increased specificity within internal cohort cross-validation and testing sets by eliminating 73.5% and 81.8% of false positives, respectively. Overall accuracy was 98.1% (95% CI: 97.4–98.8) and 99.1% (95% CI: 97.7–1.0) in cross-validation and testing sets, respectively. In the external cohort, the combined model yielded less than half as many ($n = 3$) false negative cores as the internal data set, all of which contained partial PTEN loss. The number of false positives ($n = 14$) remained stable compared to the multi-resolution technique (Tables 2 and 3). Combining the fine-tuned multi-resolution model with pixel-based classification ($n = 357$ test cores) resulted in overall accuracy of 95.24% (95% CI: 93.0–97.3) (Table 3).

Discussion

Here we demonstrate the feasibility of deep learning algorithms to automate biomarker scoring and annotation in a high-throughput TMA setting. As proof-of-principle, we deployed these algorithms to detect a well-studied tissue-based biomarker, PTEN loss in prostate cancer.

Compared to previously established qualitative analyses for protein expression, more quantitative scoring methods have the potential to provide superior molecular insights and better prognostic performance [28, 32–34]. From a clinical workflow perspective, biomarker assessment should be robust, with minimal inter- or intra-observer variability and must perform consistently across laboratories [35, 36]. All of the above prompted us to investigate the need for fully automated, standardized, cost- and time-effective approaches to biomarker assessment [36, 37].

In this study, we showed that deep learning-based algorithms can be used to effectively fully automate assessment of PTEN protein loss and annotate regions with loss in prostate cancer TMAs with accuracy ranging from 95.2 to 99.1% in two independent patient cohorts. Furthermore, deep learning-based spatial annotation of PTEN loss regions achieved favorable concordance with pathologist annotations, with median dice 0.74–0.81 across multiple testing and validation data sets. We observed the best performance when algorithms were combined. By applying a multi-resolution approach, we were able to identify the cores potentially harboring PTEN loss which was followed by a pixel-based approach for identification of specific areas of PTEN loss using IHC images. Training these cascaded models allowed us to fully automate scoring and annotation on TMA, mimicking pathologist workflow without requiring any manual annotation. We believe that similar deep learning approaches could be used for other tissue-based biomarkers to streamline sample scoring and annotation process in an unbiased, objective way in both clinical and research settings.

PTEN loss is known to be highly associated with adverse clinico-pathological outcomes at both time of diagnosis and time of surgery in prostate cancer [17, 19, 20, 29–31]. Assessment of PTEN loss has become more robust after the development of a well-validated PTEN IHC assay [20, 29–31]. Lack of clinical utilization of PTEN assessment is linked to its heterogeneous nature of expression as well as prostate cancer multifocality, which make it difficult to identify areas with PTEN loss and objectively define clinically important biomarker status on needle core biopsies [38]. In the current study, several cores containing both benign and cancer tissues from each patient were used to inherently address tissue-based as well as PTEN expression heterogeneity in multi-focal prostate cancer, and to simulate real histological scenario at the time of algorithm training.

We chose to employ a multi-resolution approach utilizing models trained at ×5, ×10, and ×20 for identification of cores potentially harboring regions of PTEN loss. We hypothesized that each resolution would balance information about tumor burden (×5/×10), architecture (×5/×10), and cellular details (×20). The multi-resolution classifier demonstrated improved performance compared to any patch-based classifier at single magnification. This is consistent with prior work by BenTaib et al. [39], who demonstrated a latent model produced from multiple magnification levels to improve subtype classification of ovarian carcinoma. Specifically, multi-resolution approach allowed to achieve higher sensitivity at the core-level compared to sensitivity at the patch-level of each resolution (×5, ×10, ×20). We have shown that combining multi-resolution classification and pixel-based spatial annotation provided the highest classification performance compared to pathologist interpretation, while maintaining high sensitivity (range: 90.6–100%) and specificity (range: 95.7–99.0%) in both cohorts.

The success of machine learning in healthcare research largely depends on proper validation of the algorithms on various external cohorts. A limited access to large data sets often leads to overfitting of algorithms to the training data sets and therefore limits the success of its direct application to external populations [40, 41]. In this study, we employed cross-validation to assess the robustness of model development. In addition, we validated a new automated detection algorithm for PTEN loss in an external independent cohort. Here we observed high performance across multiple patient splits. Cross-validated models demonstrated similar performance (AUC: 0.988–0.994) when applied to internal test set. It is worth noting that when applied to the external testing set, performance accuracy of the algorithm only reached 71.3% due to a high false positive rate (incorrect prediction of PTEN loss) when applying the same probability threshold (0.265) used in internal cohort. In addition, we observed that classification performance varied across each magnification, with ×10 maintaining superiority compared to ×5 and ×20. Reasons for the variable performance could have been due to differences in tissue processing and fixation procedures across institutions, leading to overfitting on training set. To address this limitation, we used a technique called fine-tuning [42], also referred to as transfer learning, to modify algorithms initialized from internal cohort to train a minority of samples from the external cohort in order to produce a more robust model with consistent performance across all cohorts [42]. Using minority of samples ($n = 19$ patients, 15%) from the external cohort, we were able to recover high performance with 95.2% accuracy at probability threshold 0.265 in the remaining external population ($N = 357$ cores, $n = 105$ patients) for the multi-resolution model. Future work will consider developing a “generalizable” algorithm utilizing training data from multiple institutions to apply this technique without the current dependency on fine-tuning.

Considering the heterogeneous nature of PTEN expression, previously established visual scoring criteria characterize the cases as either partial (<100% of cancer cells exhibiting PTEN loss) or complete PTEN loss (100% of cancer cells exhibiting PTEN loss). The current results demonstrated that cases with complete PTEN loss were most accurately identified by the multi-resolution algorithm. Heterogeneous (partial) PTEN loss was the primary source of false negatives (incorrectly predicted as PTEN intact) in both cohorts with rates of 4.4% (8/180) in the internal cohort and 9.4% (3/32) in the external cohort, respectively. Only one core with complete PTEN loss was misclassified by either algorithm. False positive rates for the final combined model ranged from 1.6 to 4.3% across internal and external cohorts and were observed in cores either with higher stromal content or “low PTEN” (i.e., decreased PTEN expression compared to normal cells but increased compared to threshold for loss). As “Low PTEN” cores are a common source of discordant PTEN scoring by pathologists [17], better assessment of this expression pattern will come from aligning IHC assays with orthogonal assays of PTEN status such as fluorescent in situ hybridization, FISH. Additional opportunities for improving deep learning-based approaches to PTEN assessment could come from measuring the fraction of cancer cells with PTEN loss, which has been linked to adverse prognosis [28].

This study has several important limitations. The algorithms were developed on TMA of surgical specimens which may not recapitulate tumor heterogeneity as seen in clinical samples such as needle core biopsies. As a consequence, the AI-based probability threshold of 26.5% for predicting PTEN loss developed here may need to be adjusted for different sample types and patient cohorts. Given the large imbalance between PTEN intact and PTEN loss cores, this cut-point was designed to optimize the sensitivity of multi-resolution based approach in TMA cores in cross-validation. Patch-based performance of the algorithms was observed to be lower than the multi-resolution, core-based performance. False positive classification within regions/patches of benign epithelium and cores with low cancer cell density suggests that future work should train biomarker detection algorithms within clearly separated tissue compartments (stroma vs. epithelium) from pathologist-derived annotations as ground truth within those compartments. Also, future studies should utilize cohorts with more balanced numbers of cancer and benign cores. Finally, while we demonstrated successful application of the algorithm to tissues processed and sectioned at two different institutions, fine-tuning of the model was nevertheless required. As IHC and digital scanning were performed at a single institution, the algorithm may require further modification to address variability in these procedures if done at different centers in the future studies. In addition, further correlation of AI-based approaches with clinical outcomes is warranted, which was not the scope of the current paper.

In conclusion, this work demonstrates feasibility for fully automated and robust detection and localization of PTEN loss in prostate cancer tissue samples. This novel system has great potential to streamline objective sample assessment in research and clinical laboratories, making it an unbiased and very rapid process. Such algorithms show promise to minimize subjectivity, human error and involvement, especially in resource-limited settings [4]. We expect that this fully digital workflow and robust performance will yield objective biomarker assessment and improve personalized patient care.

Code availability

All codes utilized in this study are publicly available at https://github.com/NIH-MIP/ProstateTMA_PTEN.

Acknowledgements This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This project is supported by the Prostate Cancer Foundation (Young Investigator Award). This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). Work by TV, CP, and JAS was funded by the São Paulo Research Foundation (awards #2015/22785-5, #2017/08614-9, and #2015/09111-5). Work by TJ, PGP, and DMB was supported by a Prostate Cancer Canada (PCC) award #T2014-01 proudly funded by the Movember Foundation. TJ was supported by a Transformative Pathology Fellowship awarded by the Ontario Institute for Cancer Research (OICR) through funding provided by the Government of Ontario. We would like to thank Lee R. Boudreau (Queen's Laboratory for Molecular Pathology) for tissue microarray construction and Shakeel Virk for assistance with digital imaging.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318:2199–210.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
3. Vandenberghe ME, Scott MLJ, Scorer PW, Söderberg M, Balcerzak D, Barker C. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci Rep. 2017;7:45938.
4. Chen P-HC, Gadepalli K, MacDonald R, Liu Y, Kadowaki S, Nagpal K, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat Med. 2019;25:1453–7.
5. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 2018;23:181–93.
6. Steiner DF, Macdonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42:1636–46.
7. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digit Med. 2019;2:48.
8. Kapil A, Meier A, Zuraw A, Steele KE, Rebelatto MC, Schmidt G, et al. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. Sci Rep. 2018;8:17343.
9. Khameneh FD, Razavi S, Kamasak M. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Comput Biol Med. 2019;110:164–74.
10. Saha M, Chakraborty C. Her2Net: a deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Trans Image Process. 2018;27:2189–200.
11. Brimo F, Schultz L, Epstein JI. The value of mandatory second opinion pathology review of prostate needle biopsy interpretation before radical prostatectomy. J Urol. 2010;184:126–30.
12. Rizzardi AE, Johnson AT, Vogel RI, Pambuccian SE, Henriksen J, Skubitz APN, et al. Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. Diagn Pathol. 2012;7:42.
13. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson ANA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA. 2015;313:1122–32.
14. Amin A. Understanding pitfalls of grading prostate carcinoma between biopsy and prostatectomy. AME Med J. 2018;3:24.
15. Simpson L, Parsons R. PTEN: life as a tumor suppressor. Exp Cell Res. 2001;264:29–41.
16. Álvarez-Garcia V, Tawil Y, Wise HM, Leslie NR. Mechanisms of PTEN loss in cancer: it's all about diversity. Semin Cancer Biol. 2019;59:66–79.
17. Jamaspishvili T, Berman DM, Ross AE, Scher HI, De Marzo AM, Squire JA, et al. Clinical implications of PTEN loss in prostate cancer. Nat Rev Urol. 2018;15:222–34.
18. Krohn A, Diedler T, Burkhardt L, Mayer P-S, De Silva C, Meyer-Kornblum M, et al. Genomic deletion of PTEN is associated with tumor progression and early PSA recurrence in ERG fusion-positive and fusion-negative prostate cancer. Am J Pathol. 2012;181:401–12.
19. Troyer DA, Jamaspishvili T, Wei W, Feng Z, Good J, Hawley S, et al. A multicenter study shows PTEN deletion is strongly associated with seminal vesicle involvement and extracapsular extension in localized prostate cancer. Prostate. 2015;75:1206–15.
20. Lotan TL, Wei W, Ludkovski O, Morais CL, Guedes LB, Jamaspishvili T, et al. Analytic validation of a clinical-grade PTEN immunohistochemistry assay in prostate cancer by comparison with PTEN FISH. Mod Pathol. 2016;29:904–14.
21. Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part I: risk stratification, shared decision making, and care options. J Urol. 2018;199:683–90.
22. National Comprehensive Cancer Network. Prostate cancer. https://www.nccn.org/store/login/login.aspx?ReturnURL=

https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf (2019).

23. McKenney JK, Simko J, Bonham M, True LD, Troyer D, Hawley S, et al. The potential impact of reproducibility of gleason grading in men with early stage prostate cancer managed by active surveillance: a multi-institutional study. J Urol. 2011;186:465–9.
24. Sedelaar JPM, Schalken JA. The need for a personalized approach for prostate cancer management. BMC Med. 2015;13:109.
25. Eggener SE, Rumble RB, Armstrong AJ, Morgan TM, Crispino T, Cornford P, et al. Molecular biomarkers in localized prostate cancer: ASCO guideline. J Clin Oncol. 2020;38:1474–94.
26. Lotan TL, Gurel B, Sutcliffe S, Esopi D, Liu W, Xu J, et al. PTEN protein loss by immunostaining: analytic validation and prognostic indicator for a high risk surgical cohort of prostate cancer patients. Clin Cancer Res. 2011;17:6563–73.
27. Lokman U, Erickson AM, Vasarainen H, Rannikko AS, Mirtti T. PTEN loss but not ERG expression in diagnostic biopsies is associated with increased risk of progression and adverse surgical findings in men with prostate cancer on active surveillance. Eur Urol Focus. 2017;4:867–73.
28. Jamaspishvili T, Patel PG, Niu Y, Vidotto T, Caven I, Livergant R, et al. Risk stratification of prostate cancer through quantitative assessment of PTEN loss (qPTEN). J Natl Cancer Inst. 2020; online ahead of print.
29. Lotan TL, Gurel B, Sutcliffe S, Esopi D, Liu W, Xu J, et al. PTEN protein loss by immunostaining: analytic validation and prognostic indicator for a high risk surgical cohort of prostate cancer patients. Clin Cancer Res. 2011;17:6563–73.
30. Ahearn TU, Pettersson A, Ebot EM, Gerke T, Graff RE, Morais CL, et al. A prospective investigation of PTEN loss and ERG expression in lethal prostate cancer. J Natl Cancer Inst. 2015;108: djv346.
31. Lotan TL, Carvalho FLF, Peskoe SB, Hicks JL, Good J, Fedor HL, et al. PTEN loss is associated with upgrading of prostate cancer from biopsy to radical prostatectomy. Mod Pathol. 2015;28:128–37.
32. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILS) in breast cancer: recommendations by an International TILS Working Group 2014. Ann Oncol. 2015;26:259–71.
33. Stålhammar G, Fuentes Martinez N, Lippert M, Tobin NP, Mølholm I, Kis L, et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. Mod Pathol. 2016;29:318–29.
34. Cheang MCU, Martin M, Nielsen TO, Prat A, Voduc D, Rodriguez-Lescure A, et al. Defining breast cancer intrinsic subtypes by quantitative receptor expression. Oncologist. 2015;20:474–82.
35. Duffy MJ, Sturgeon CM, Sölétormos G, Barak V, Molina R, Hayes DF, et al. Validation of new cancer biomarkers: a position statement from the european group on tumor markers. Clin Chem. 2015;61:809–20.
36. Selleck MJ, Senthil M, Wall NR. Making meaningful clinical use of biomarkers. Biomark Insights. 2017;12:1177271917715236.
37. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. npj Precis Oncol. 2017;1. https://doi.org/10.1038/s41698-017-0022-1.
38. Tosoian JJ, Guedes LB, Morais CL, Mamawala M, Ross AE, De Marzo AM, et al. PTEN status assessment in the Johns Hopkins active surveillance cohort. Prostate Cancer Prostatic Dis. 2019;22:176–81.
39. BenTaieb A, Li-Chang H, Huntsman D, Hamarneh G. A structured latent model for ovarian carcinoma subtyping from histopathology slides. Med Image Anal. 2017;39:194–205.
40. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15:e1002683.
41. Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318: 2211–23.
42. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc. 2018;25:945–54.