Title

Stress testing deep learning models for prostate cancer detection on biopsies and surgical specimens.

Permalink

https://escholarship.org/uc/item/6w53b300

Journal

The Journal of pathology and bacteriology, 265(2)

Authors

Flannery, Brennan
Sandler, Howard
Lal, Priti
et al.

Publication Date

2025-02-01

DOI

10.1002/path.6373

Peer reviewed

Journal of Pathology
J Pathol February 2025; **265:** 146–157
Published online 11 December 2024 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/path.6373

ORIGINAL ARTICLE

Stress testing deep learning models for prostate cancer detection on biopsies and surgical specimens

Brennan T Flannery[1], Howard M Sandler[2], Priti Lal[3], Michael D Feldman[4], Juan C Santa-Rosario[5], Tilak Pathak[6], Tuomas Mirtti[6], Xavier Farre[7], Rohann Correa[8], Susan Chafe[9], Amit Shah[10], Jason A Efstathiou[11], Karen Hoffman[12], Mark A Hallman[13], Michael Straza[14], Richard Jordan[15], Stephanie L Pugh[16], Felix Feng[17] and Anant Madabhushi[6,18*]

[1] *Case Western Reserve University, Cleveland, OH, USA*
[2] *Cedars-Sinai Medical Center, Los Angeles, CA, USA*
[3] *University of Pennsylvania, Philadelphia, PA, USA*
[4] *Indiana University, Indianapolis, IN, USA*
[5] *CorePlus, Carolina, Puerto Rico*
[6] *Emory Winship Cancer Institute, Atlanta, GA, USA*
[7] *Public Health Agency of Catalonia, Catalonia, Spain*
[8] *London Health Science Centre, London, ON, Canada*
[9] *Cross Cancer Institute, Edmonton, AB, Canada*
[10] *WellSpan Health, York, PA, USA*
[11] *Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*
[12] *The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[13] *Fox Chase Cancer Center, Philadelphia, PA, USA*
[14] *Medical College of Wisconsin, Milwaukee, WI, USA*
[15] *NRG Oncology Biospecimen Bank, San Francisco, CA, USA*
[16] *NRG Oncology Statistics and Data Management Center, Philadelphia, PA, USA*
[17] *University of California San Francisco, San Francisco, CA, USA*
[18] *Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA*

**Correspondence to: A Madabhushi, Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA. E-mail: anantm@emory.edu*

Abstract

The presence, location, and extent of prostate cancer is assessed by pathologists using H&E-stained tissue slides. Machine learning approaches can accomplish these tasks for both biopsies and radical prostatectomies. Deep learning approaches using convolutional neural networks (CNNs) have been shown to identify cancer in pathologic slides, some securing regulatory approval for clinical use. However, differences in sample processing can subtly alter the morphology between sample types, making it unclear whether deep learning algorithms will consistently work on both types of slide images. Our goal was to investigate whether morphological differences between sample types affected the performance of biopsy-trained cancer detection CNN models when applied to radical prostatectomies and vice versa using multiple cohorts (N = 1,000). Radical prostatectomies (N = 100) and biopsies (N = 50) were acquired from The University of Pennsylvania to train (80%) and validate (20%) a DenseNet CNN for biopsies (M^B), radical prostatectomies (M^R), and a combined dataset (M^{B+R}). On a tile level, M^B and M^R achieved F1 scores greater than 0.88 when applied to their own sample type but less than 0.65 when applied across sample types. On a whole-slide level, models achieved significantly better performance on their own sample type compared to the alternative model ($p < 0.05$) for all metrics. This was confirmed by external validation using digitized biopsy slide images from a clinical trial [NRG Radiation Therapy Oncology Group (RTOG)] (NRG/RTOG 0521, N = 750) via both qualitative and quantitative analyses ($p < 0.05$). A comprehensive review of model outputs revealed morphologically driven decision making that adversely affected model performance. M^B appeared to be challenged with the analysis of open gland structures, whereas M^R appeared to be challenged with closed gland structures, indicating potential morphological variation between the training sets. These findings suggest that differences in morphology and heterogeneity necessitate the need for more tailored, sample-specific (i.e. biopsy and surgical) machine learning models.

Keywords: deep learning; machine learning; convolutional neural networks; morphology; prostate cancer; biopsy; radical prostatectomy; generalizability; interpretability

Received 22 May 2024; Revised 21 October 2024; Accepted 23 October 2024

Conflict of interest statement: JAE declares in the last 36 months consulting fees from Blue Earth Diagnostics, Boston Scientific, AstraZeneca, Genentech, Astellas/Pfizer, Lantheus, payment or honoraria for lectures, presentations, speakers, bureaus, manuscript writing, or educational events from IBA, Elekta, and UpToDate, participation on a data safety monitoring board or advisory board at Merck, Roivant Pharma, Myovant Sciences, Janssen, Bayer Healthcare, Progenics Pharmaceuticals, Pfizer, Astellas, Gilead, Lantheus, Blue Earth Diagnostics, Angiodynamics, leadership or fiduciary

role in other board, society, committee, or advocacy group, paid or unpaid work for NRG Vice-Chair GU steering committee; board member (unpaid): Massachusetts Prostate Cancer Coalition (MPCC), American College of Radiation Oncology (ACRO), and Radiation Oncology Institute (ROI). KH declares in the last 36 months research funding from Varian Medical Systems and Janssen, participation on a data safety monitoring board or advisory board at MD Anderson DSMB (ended 2022), leadership or fiduciary role in other board, society, committee, or advocacy group, and paid or unpaid work as ASTRO representative to the Commission on Cancer. AM declares support for the work reported in this manuscript without time limit as an equity holder in Picture Health, Elucid Bioimaging, and Inspirata Inc. Currently he serves on the advisory board of Picture Health and SimBioSys. He also currently consults for SimBioSys. He also has sponsored research agreements with AstraZeneca and Bristol Myers-Squibb. His technology has been licensed to Picture Health and Elucid Bioimaging. He is also involved in three different R01 grants with Inspirata Inc. TM declares support for the work reported in this manuscript without time limit from the Academy of Finland with a grant to the University of Helsinki and declares in the last 36 months a grant for a research scholarship at Emory University from the Sigrid Juselius Foundation. HMS declares support for the work reported in this manuscript without time limit prior funding from ACR/NRG Oncology for role as chair of GU Cancer Committee, declares in the last 36 months leadership or fiduciary role in other board, society, committee, or advocacy group, and paid or unpaid work on the ASTRO board of directors. No other conflicts of interest were declared.

Introduction

Nearly one-third of all male cancer diagnoses are of prostate cancer, making it the second leading cause of cancer death among men. [1]. Hematoxylin and eosin (H&E)-stained core needle biopsies (core needle BXs) are the main method for diagnosing prostate cancer [2]. Samples are acquired by puncturing the prostate with a core needle device and extracting a tissue sample that is subsequently processed and placed on a slide for pathologist review [3]. If cancer is observed, it is graded by a pathologist to determine the extent and aggressiveness of the disease [4], which then influences the subsequent course of treatment, e.g. active surveillance versus radical prostatectomy (RP). For those patients that opt for a surgical intervention typically via RP, the individual tissue specimens are sectioned (either quartered or whole) and the sections mounted on slides [5]. Cancer within these samples is graded to gauge the risk of post-treatment adverse patient outcomes such as biochemical recurrence, metastasis, or death. BXs and RPs serve different roles in the clinical process and have differences in preparation that are known to impact observed cancer heterogeneity [6] and histopathologic characteristics. [7].

Automated methods using artificial intelligence (AI) have been developed to evaluate prostate cancer tissue samples more quickly and with less bias than manual investigation [8]. AI models have been developed for slide-level classification [9–11] and prediction [12,13], as well as for smaller subsections of slides referred to as tile-level cancer detection [10,11,14]. While some of these approaches have been clinically validated [9,10] and have demonstrated high accuracy in BXs, very few of them have been tested on RPs. This includes FDA-approved tools such as that of Paige AI, which was developed to identify prostate cancer on core needle BXs[15]. Though extensively validated on BXs [15–18], it is unclear whether similar models would maintain consistent performance when applied to RPs. Such models could potentially perform well across sample types given both BXs and RPs contain histologically prepared sections of prostate and tumor tissue prepared with the same stain (H&E) and digitally scanned in the same manner. To our knowledge, there has not been a dedicated study focusing specifically on understanding whether morphologic differences exist between RPs [19] and BXs [8,9,14,18] and how those differences affect AI models trained for cancer detection.

The differences in acquisition, preparation, and morphology between RPs and BXs make it unclear whether cancer detection models trained on one sample type will translate to the other. With the current rise in FDA-approved AI-based cancer detection technologies [8–10,19], there is a need for clarity as to which types of samples these models can be applied. One of the primary concerns when attempting to translate models between sample types is the presence of batch effects [20] and addressing intersite differences. Batch effects in training AI models can be addressed through data augmentation during training, which prevents models from learning dataset-specific features that disrupt generalizability [21]. More specifically for H&E images, stain normalization approaches have been developed to reduce differences between datasets that arise from different tissue staining procedures, environments, or imaging systems [22]. It is highly likely that once these factors are taken into account, performance differences between models will be driven primarily by morphological differences between sample types.

The aim of this study was to investigate whether preparation differences between RPs and BXs resulted in morphological differences that drive differential performance in cancer detection models when RP deep learning models are applied to BXs and vice versa. The scientific premise of this work was that the BX model would detect cancer more accurately when applied to BX data compared to RP data. We also expected that the RP model would detect cancer more accurately when applied to RP data compared to BX data. Our goal was to determine whether morphologically driven performance differences are present in cancer detection models.

The data used in this study comprised HE-stained slides corresponding to 100 RPs and 50 BXs from The University of Pennsylvania. An additional set of 98 BXs from CorePlus Servicios Clínicos y Patológicos was used as an independent, qualitative external validation set. Another additional set of 750 BXs from the NRG Radiation Therapy Oncology Group (NRG/RTOG) 0521 clinical trial was used for both qualitative and quantitative external validation. DenseNet [23] models were trained for both RPs and BXs, which were subsequently tested on both sample types and on an independent set of BXs acquired from CorePlus.

Materials and methods

Study design

This study used 1,000 patient-related samples from four cohorts and multiple institutions (Table 1, Figure 1A). Digitally scanned images from RP HE-stained slides at magnification ×40 were acquired from The University of Pennsylvania for 100 patients (S^R). Digitally scanned core needle BXs were acquired from the same institution at a magnification of ×40 for a different set of 50 patients (S^B). Both cohorts were annotated by expert pathologists for cancerous regions at variable magnifications (changed as needed by the pathologist). Slide-level Gleason score and Gleason grade information was available for all patients (Figure 2). Since the distributions of Gleason grades were relatively similar between cohorts, with the exception of some high-grade cancer in RP specimens, these cohorts were designated for model training. Additionally, these cohorts were chosen because they were all digitally scanned using the same device (Hamamatsu NanoZoomer S360, Hamamatsu Photonics, Bridgewater, NJ, USA), reducing scanning-related differences between the datasets. The scanners used to digitize other datasets available to us either did not match these datasets or could not be verified. An external validation set of unannotated biopsy samples was acquired from CorePlus Servicios Clínicos y Patológicos for 100 patients (S^C). An additional multi-institutional external validation set of 750 biopsy samples from 350 patients was acquired from a subset of the NRG/RTOG 0521 clinical trial [24], of which 28 patients had cancer annotations (S^{RTOG}). All images underwent quality testing using HistoQC [20], an open-source quality control tool for digital pathology slides, which indicated most slides were usable for model training or validation (Figures 1B and 3). All slides were downsampled to magnifications of ×1, ×5, and ×10 to decrease memory burden and increase training speed. Each slide was split into 256 × 256 pixel tiles that were assigned a binary designation of cancer (greater than 30% cancer present) or noncancer (less than 30% cancer) based on the pathologist annotation. Image tiles containing a relatively small amount of tissue (less than 20%) were discarded. All data were acquired with approval of the respective hospital Institutional Review Board, with informed consent waived due to retrospective collection.

Deep learning models for prostate cancer detection

All models developed in this study utilized the DenseNet architecture [23] (Figure 1C). This is a convolutional network where each layer is densely connected to every following layer. This structure prevents overfitting, strengthens feature propagation, and reduces the vanishing gradient problem that plagues large neural networks. The input to these models were the extracted tiles and the output a vector containing the probability of the tile belonging to each class. All models were developed in Python using PyTorch (version 1.12 with CUDA compatibility, The Linux Foundation, San Diego, CA, USA).

This study comprised the following three experiments. (1) A DenseNet model was trained for BXs using S^B (M^B) and RPs using S^R (M^R) separately. They were then evaluated on their respective internal validation sets ($S^{B,V}$, $S^{R,V}$). This process was repeated at magnifications ×1, ×5, and ×10. (2) M^B was evaluated at all magnifications on S^B while M^R was evaluated across all magnifications on S^R. Both models were then evaluated on S^C and S^{RTOG}. (3) A combined model was trained using both S^B and S^R at magnification ×5 (M^{B+R}) and validated on a combination of $S^{B,V}$ and $S^{R,V}$. Additionally, experiments 1 and 2 were repeated for magnification ×5 for three additional neural network architectures (ResNet, EfficientNet, ResNext).

Mitigation of batch effects

Once batch effects are considered and reduced, any differences in model performance were expected to likely be engendered by morphological differences between BXs and RPs. Four approaches were taken to reduce batch effects between S^B and S^R. (1) The datasets were acquired from the same institution to normalize the slide preparation and digitization protocols. (2) Both datasets were investigated using Uniform Manifold Approximation and Projection (UMAP) embeddings [25] to ensure there were no differences in clustering between the two types of samples prior to training. (3) Random image augmentation was applied during model training [21]. These strategies included random cropping, hue saturation value (HSV) shifting, contrast adjustment, brightness adjustment, and rotations. The probability that any one of these adjustments would occur was 50%. (4) Stain normalization using Macenko normalization [26] was applied

Table 1. All datasets used in this study. We utilized 1,000 slides from 600 patients across three institutions.

Institution	Short name	Sample type	No. samples	No. patients
UPenn biopsies	S^B	Biopsy	50	50
UPenn radical prostatectomies	S^R	Radical prostatectomy	100	100
CorePlus	S^C	Biopsy	100	100
NRG/RTOG 0521	S^{RTOG}	Biopsy	750	350
		Total	1,000	600

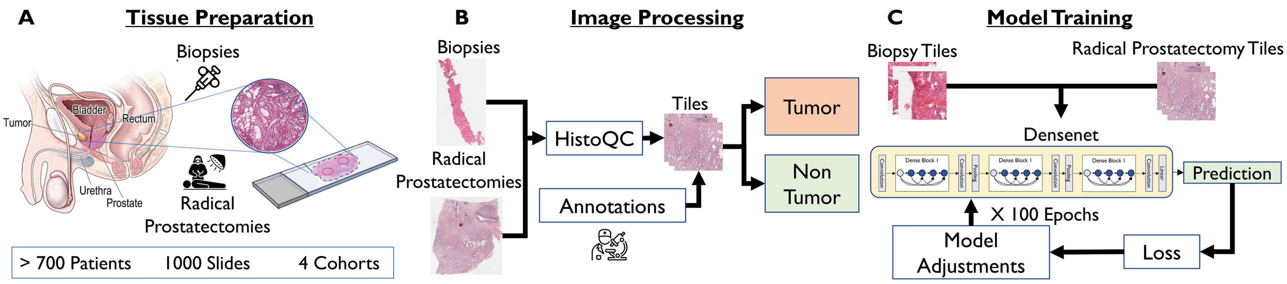

Figure 1. Demonstration of experimental workflow in this study. (A) Data preparation from both biopsies and radical prostatectomies. (B) Image processing to generate tiles with tumor and nontumor classifications. (C) Iterative model training process for Densenet models. Partially created with BioRender.com.

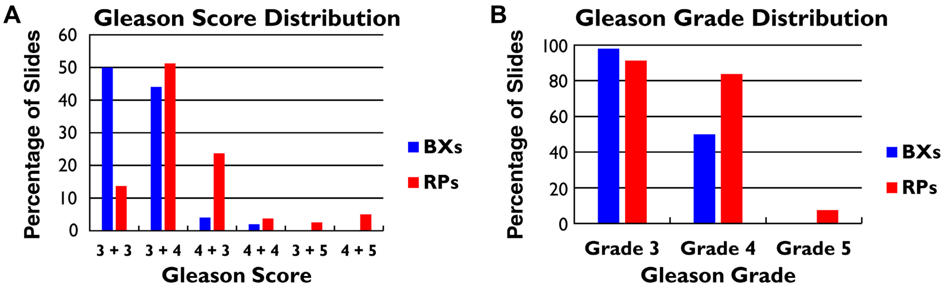

Figure 2. Distribution of pathological scoring in the two training cohorts. (A) Gleason score distributions in both UPenn datasets. (B) Gleason grade distributions in both UPenn datasets.

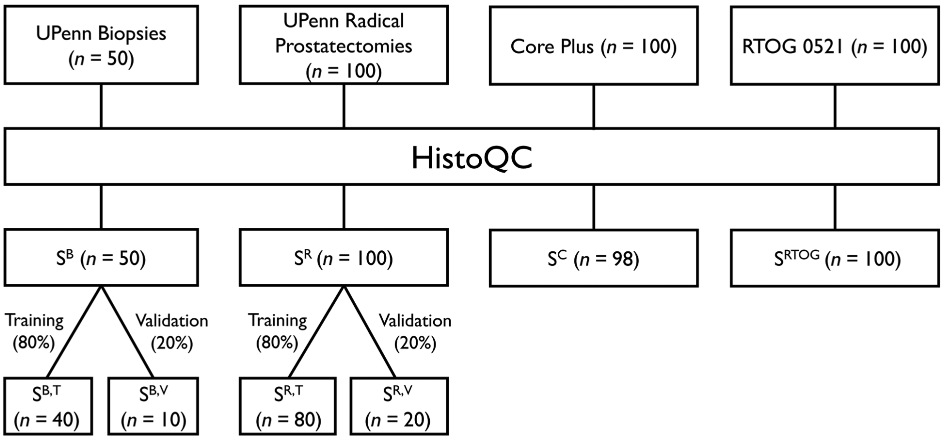

Figure 3. CONSORT diagram of different cohorts used in this study. Only two patients were removed during quality testing using HistoQC.

during training using a randomly chosen tile from a balanced subset of the data. Every tile within a batch was equally likely to be normalized to a RP tile as it was to a BX tile, reducing model reliance on stain differences (supplementary material, Figure S1).

Strategies to mitigate differences in annotation were also considered. Differences between human annotators can result in different ground truth labels when training machine learning models, thereby affecting model performance [27]. Semantic segmentation models are particularly vulnerable to this because they make predictions at a very small scale (per pixel). However, tile-wise classification helps minimize the effect of annotation differences by considering a region of tissue when performing the classification task. All models trained in this study were tile-based classification models, intentionally designed to take advantage of this mitigation strategy.

Statistical evaluation

Each training dataset (S^B, S^R) was split into training (80%) and internal validation (20%) subsets, with each model being trained for 100 epochs with a binary cross entropy loss function [28] and an Adam optimizer [29] (lr $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). The model that achieved the smallest loss on the training set was saved for further evaluation. Experiments 1 and 2 were conducted from the perspective of both tiles and whole slides. Tile-level performance was evaluated using sensitivity, specificity, and F1 score. Whole-slide performance was evaluated using F1 score, area under the receiver operating curve (AUC), and Dice score. The F1 score combines precision and recall, or a weighted ratio of true positives to false positives (FPs) and false negatives (FNs), to create a comprehensive accuracy metric for classification problems. The Dice score is a measure of overlap between an annotated region and the predicted cancer region from the model. Slide-level performance was computed for each patient using both M^B and M^R and compared using *t*-tests ($\alpha = 0.05$). These comparisons were also computed using S^{RTOG}.

Because cancer annotations were not available for S^C and many of the slide images in S^{RTOG}, an additional qualitative comparison was performed. M^B and M^R were applied to each BX slide in its entirety to generate segmentation overlays that were subsequently presented to two expert pathologists (TP, TM), who in turn evaluated model performance. All slides were stain normalized prior to generating the overlays. Each reviewer was presented with segmentation overlays from both models; however, they were blinded to which model each overlay belonged to. For each slide image, both reviewers indicated which model they believed was better at cancer detection. This determination was made based on the amount of correctly identified cancer tissue (true positives) as well as the amount of incorrectly classified benign tissue (FPs). Once the evaluation for all patients was complete, a thorough review was conducted on cases for which the reviewers disagreed. For each of these cases, a consensus decision was reached between the two reviewers as to which model resulted in superior cancer detection. The total number of slides for which each model was preferred by the reviewers was compared. To have sufficient power in our analysis (80%) to detect the difference in model preference, we compared 60 randomly selected slides from S^{RTOG}.

Additionally, M^B and M^R were applied to S^C at magnification ×5. The agreement between the classifications was determined by applying the model across the entire slide and then calculating the Dice score between the labeled regions.

Interpretability of deep learning models

Once cancer detection models were trained for both BXs and RPs, we implemented two approaches to assess model interpretability. (1) Grad-cam activation maps [30] were acquired for both M^R and M^B on both S^B and S^R. These maps are visual representations of the locations in the image that the model used to make classification decisions based on the strength of the model gradients. The activation maps were qualitatively evaluated by a nonexpert (BF) and observations were confirmed by an expert pathologist (TP or TM). The evaluation included presenting expert pathologists with 5× magnification tiles as well as the model activation maps associated with those tiles. The pathologists were tasked with identifying structures (glands) that were highlighted in model attention maps and identifying general characteristics of those structures (size, lumen, cancerous or benign). (2) All FP and FN tiles were examined by the same two experts. Observations on tile appearance (including gland shape and size) and morphology from both pathologists were compiled to create a comprehensive consensus review of model performance. An additional quantitative review was performed to confirm pathologist observations. All tiles from $S^{B,V}$, along with an equal sized randomly chosen subset of tiles from $S^{R,V}$, were reviewed by three expert pathologists (TP, TM, XF). Each tile was graded (Gleason grade 3–5), and any observed variant morphology was noted. The pathologists reviewed difficult-to-grade tiles and created a consensus grade for each. χ^2 tests ($\alpha = 0.05$) were conducted to compare the number of true positives (TPs), true negatives (TNs), FPs and FNs between M^B and M^R for all grade classifications within both RPs and BXs.

Results

Experiment 1: validating surgical and biopsy-specific cancer detection models on surgical and biopsy images respectively

Both M^B and M^R achieved F1 scores greater than 0.89 on $S^{B,V}$ and $S^{R,V}$, respectively. On a tile level, M^B had an F1 score of 0.93, with sensitivity of 0.92 and specificity of 0.68 (Table 2) at a magnification of ×5 on the $S^{B,V}$. M^B achieved similar performance but lower metrics using ×1 and ×10 tiles. Similarly, M^R had an F1 score of 0.89, with a sensitivity of 0.84 and specificity of 0.97 at ×5 magnification on the $S^{R,V}$. M^R achieved very similar performance on ×1 and ×10 tiles.

Experiment 2: validating surgical and biopsy-specific cancer detection models on different sample types

Both M^B and M^R achieved lower performance metrics when applied to $S^{R,V}$ and $S^{B,V}$, respectively. On a tile level at ×5 magnification, M^B applied to the $S^{R,V}$ set resulted in a F1 score of 0.64, with a sensitivity of 0.49 and a specificity of 0.91 (Table 2). M^R demonstrated a similar decrease in performance when applied to $S^{B,V}$ at ×5 magnification, achieving an F1 score of 0.53, a sensitivity of 0.94, and a specificity of 0.23. These trends were consistent for both ×1 and ×10 magnifications. Similar trends were observed when M^B and M^R were trained using ResNet, EfficientNet, and ResNext architectures (supplementary material, Table S1). M^B consistently outperforms M^R on BXs in terms of F1 score regardless of architecture, while M^R consistently outperforms M^B on RPs for all architectures. F1 score, sensitivity, and specificity were lower for these architectures compared to DenseNet.

Slide-level performance reflected similar trends as well (Table 3). M^B achieved higher F1, AUC, and Dice scores compared to M^R on both the S^B as well as S^{RTOG}.

Table 2. Tile-level performance metrics for cancer detection models trained using tiles at ×1, ×5, and ×10 magnification with various internal validation sets from S^B and S^R. Datasets include biopsies (BX), radical prostatectomy (RP), and a RP + BX set. Metrics include sensitivity (Se), specificity (Sp), and F1 score (F1).

	M^B			M^R			M^{B+R}		
	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
Validation Set					×5				
BX	0.92	0.68	0.93	0.94	0.23	0.53	0.96	0.41	0.83
RP	0.49	0.91	0.64	0.85	0.94	0.88	0.84	0.97	0.89
RP + BX	0.66	0.83	0.75	0.87	0.75	0.77	0.87	0.95	0.91
					×10				
BX	0.93	0.59	0.85	0.88	0.37	0.6	0.92	0.39	0.83
RP	0.72	0.66	0.21	0.84	0.95	0.88	0.87	0.90	0.87
RP + BX	0.75	0.64	0.62	0.86	0.72	0.78	0.87	0.89	0.88
					×1				
BX	0.89	0.59	0.81	0.92	0.36	0.36	0.93	0.42	0.84
RP	0.77	0.66	0.11	0.89	0.97	0.91	0.87	0.90	0.87
RP + BX	0.78	0.61	0.66	0.89	0.73	0.73	0.87	0.89	0.88

Table 3. Slide-level performance metrics for M^B and M^R. Metrics include F1 score (F1), area under the receiver operating curve (AUC), and dice score (Dice). Metrics for both UPenn datasets and NRG/RTOG 0521 are shown.

Dataset	M^B			M^R		
	F1	AUC	Dice	F1	AUC	Dice
BX	0.48 ± 0.27	0.70 ± 0.13	0.53 ± 0.23	0.35 ± 0.28	0.62 ± 0.13	0.36 ± 0.27
RP	0.17 ± 0.20	0.48 ± 0.07	0.17 ± 0.20	0.69 ± 0.19	0.66 ± 0.11	0.69 ± 0.14
RTOG	0.46 ± 0.36	0.70 ± 0.17	0.57 ± 0.25	0.23 ± 0.24	0.58 ± 0.09	0.39 ± 0.16

Slide-level performance from M^B was relatively consistent across both datasets.

Using *t*-tests to compare the distributions of each metric between M^R and M^B for both S^B and S^{RTOG} demonstrated that M^B achieved significantly higher F1 scores, AUC values, and Dice scores (Figure 4). M^R achieved higher accuracy metrics on the S^R compared to M^B. The distribution of M^R performance metrics compared to M^B was also significantly higher based on *t*-tests. These differences in model predictions were observed qualitatively on S^C as well. Tile level overlap between M^B and M^R was 0.29 based on Dice score, indicating very little agreement between the models.

Qualitative analysis of model outputs on S^{RTOG} revealed that M^B performed much better than M^R, with M^B detecting much more cancer tissue compared to M^R without FP classifications (Figure 5E,F). M^B was preferred by expert pathologists for 77% of slides. When considering only slides that had a minimal acceptable amount of cancerous tissue to be usable in a downstream analysis, M^B was preferred in 92% of slides.

Experiment 3: validating a combined surgical and biopsy image-trained model

The combined RP and BX model M^{B+R} demonstrated high performance in sensitivity, specificity, and F1 score for the combined validation set as well as $S^{R,V}$ (Table 2). While achieving a 0.83 F1 score on the UPenn BX internal validation set, it also had a lower specificity (0.41) compared to M^R and M^B.

Model interpretability

When applied to RPs, M^B mistook many large cancerous glandular structures as benign (Figure 5A(i,ii)), meaning it had many FN errors when the glands in question were large. However, M^B consistently identified smaller cancerous glands as cancer. Similarly, M^B mistook some large benign glandular structures as cancerous (Figure 5B(i,ii)). M^R had very few FNs when applied to RPs, but it had many FPs where it identified benign glandular structures as cancer. The behavior of both models was consistent when applied to BXs. M^B resulted in similar errors, struggling to identify cancer presenting as very large glandular structures (Figure 5A (iii,iv)). However, it had very few FP, only mistaking some areas along the edge of the sample as cancer (Figure 5B(iii,iv)). On BXs, M^R identified most glands with visible lumen as cancer regardless of their true classification (Figure 5D(iii,iv)) while struggling to identify smaller, closed glands (Figure 5C(iii,iv)). On RPs, M^R misclassified a small number of border tiles, small glands, and prostatic stones (Figure 5C(i,ii), 5D (i,ii)).

Distinct differences were observed between activation maps for M^B and M^R (Figure 6). M^B shows much more attention to closed glands while avoiding open glands. This is evident primarily on BXs (Figure 6A,B) but can still be seen on RPs (Figure 6C,D). M^R exhibits the opposite behavior, showing more attention to open glands (with lumen) compared to closed glands. This can be observed in both RPs and BXs.

Quantitative evaluation of the influence of cancer grade and variant morphology on model predictions

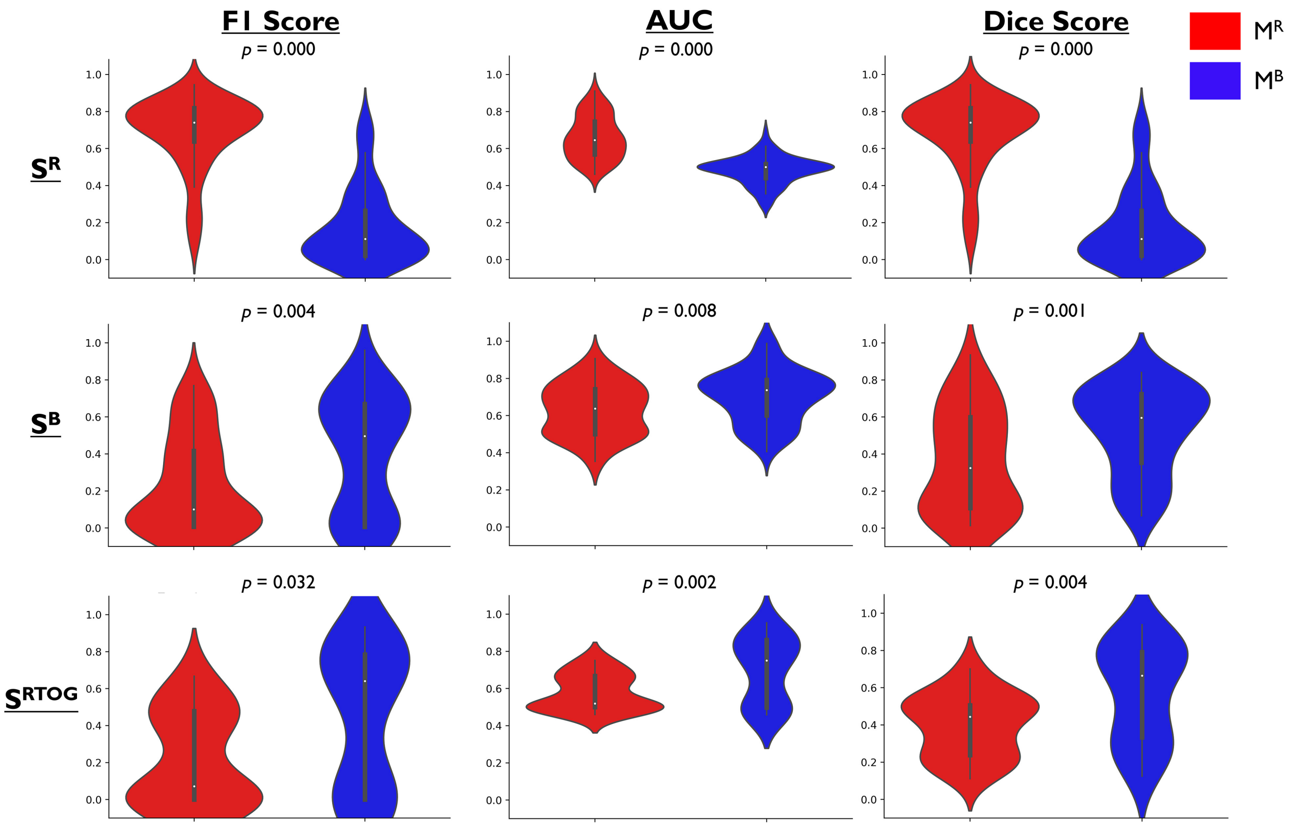

Figure 4. Slide-level performance of both M^R and M^B on all datasets. All comparisons were significantly different ($p < 0.05$). M^B obtained significantly better performance on both the UPenn BX dataset and NRG/RTOG 0521, which also contains biopsies exclusively. M^R obtained significantly better performance on UPenn RP dataset.

revealed significant differences between M^R and M^B. M^B had significantly more TNs and fewer FPs in benign tiles compared to M^R in both $S^{B,V}$ and $S^{R,V}$ (supplementary material, Figure S3A,D). M^R had significantly more TPs and fewer FNs in $S^{R,V}$ for both grades 3 and 4 cancers (supplementary material, Figure S3E,F). Neither model had no significant differences in the amount of TPs or FNs in $S^{B,V}$ for both grades 3 and 4 cancers (supplementary material, Figure S3B,C). No Gleason grade 5 tiles were observed in $S^{B,V}$ or $S^{R,V}$. Comparison of the presence of variant morphologies within TP, FN, TN, and FP tiles reveals some distinct differences between M^R and M^B (supplementary material, Table S2). M^B mistakes more cribriform tiles for benign compared to M^R. M^R has more FN classifications of tiles containing foamy glands compared to M^B. Lastly, M^B has more FN classifications of ductal cancer tiles compared to M^R.

Discussion

This study investigated whether prostate cancer detection AI models trained using BXs were generalizable to RPs and vice versa. Additionally, we investigated whether there were morphologically driven performance differences between models trained on different sample types (RPs versus BXs). To accomplish this, differences in data collection that led to batch effects were minimized, allowing for accentuation of the inherent morphological differences on account of preparation differences between the two different sample types. This was accomplished using samples from the same institution, investigating UMAP embeddings (supplementary material, Figure S2), using both image augmentations and stain normalization [26] during model training, and developing tile-based classification models.

While this study does not claim to have uncovered a definitive reason for the differences observed between AI models across different sample types, it seems highly likely that the reason is on account of the heterogeneity of prostate tissue morphology between RPs and BX samples [6]. Comprehensive review of the FP and FN errors of each AI model revealed that they both appeared to be impacted by the distinct morphological patterns across sample types (Figure 5). M^B produced more FN errors, particularly when presented with large glands on RPs. This suggests that M^B struggled to accurately classify large glands on RPs, indicating that these structures may not occur often within BXs. A potential explanation is that BX samples are limited in width by the width of the BX needle (about 1.0 mm), which does not allow for adequate representation of larger glands or groups of glands whose size could exceed this width. These structures are captured well on RPs because there is no such

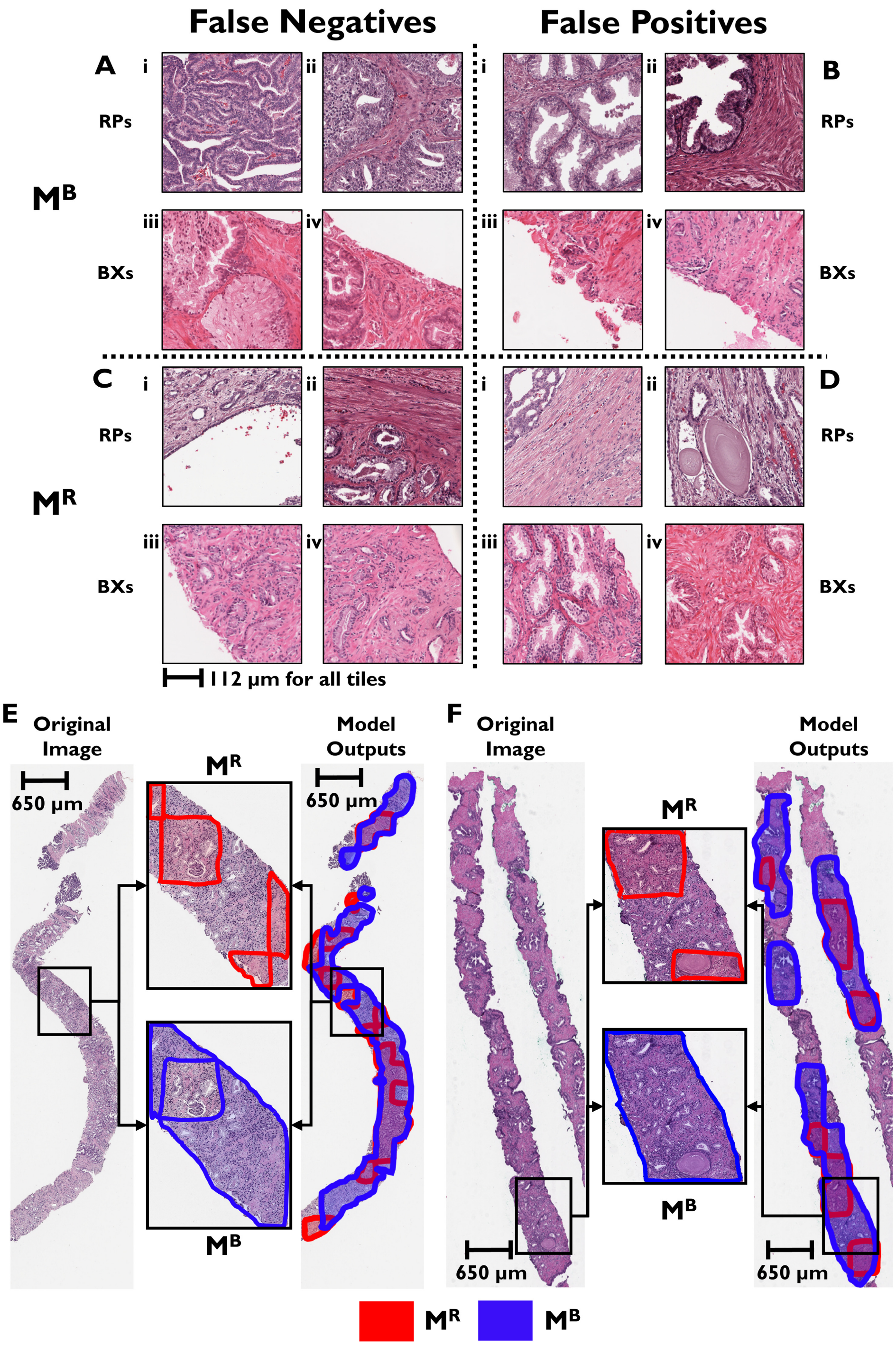

Figure 5. Example outputs from M^B and M^R. (A–D) Examples of incorrect classifications from both models on both RPs and BXs. (A and B) M^B struggles to correctly classify large glandular structures and results in errors along sample edges. (C and D) M^R struggles to correctly classify smaller glandular structures, particularly on biopsies, and along the edges of tissue. M^R also struggles with the relatively rare occurrence of prostatic stones. (E and F) Overlays for model predictions on two examples from NRG/RTOG 0521. M^R predictions are indicated in red, while M^B predictions are indicated in blue.

limitation to the size of the sample, allowing full representation of larger structures. Since M^R performed very well on RPs, there were very few FNs and FPs. Of these, M^R only struggled with the borders of annotated regions that contained both cancerous and benign tissue, the borders of the sample, or tiles that contained unique structures such as prostatic concretions, which are large groups of aggregated proteins [31]. M^R had many FPs when applied to BXs, mostly struggling with smaller glands, while M^B struggled with larger glands (Figure 5). This might reflect differences in glandular appearance on BXs and RPs, in turn impacting the predictions of cancer detection models.

The differences in model performance on account of morphological differences between samples is also supported by the different activation patterns of M^R

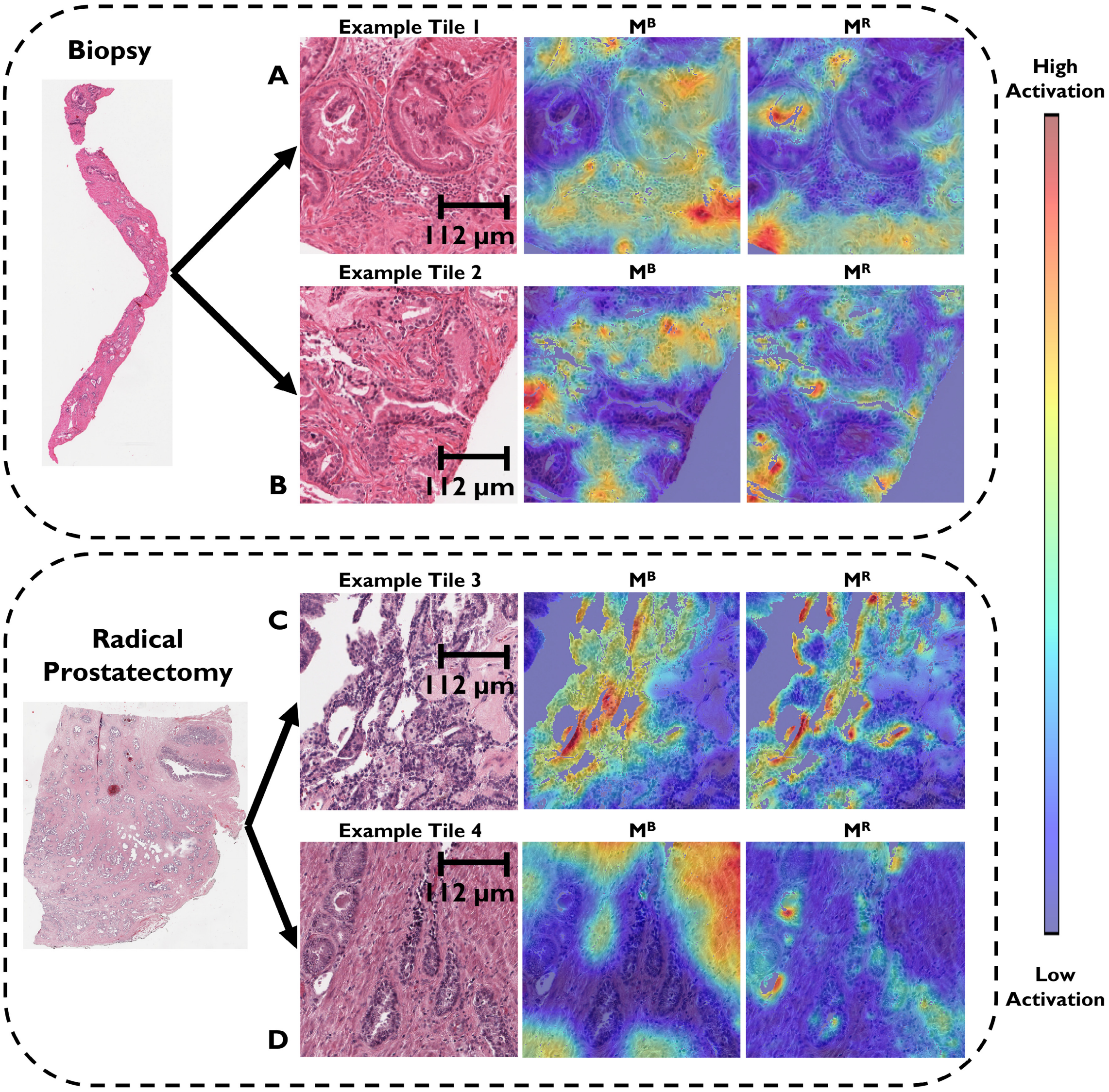

Figure 6. Grad-Cam attention maps for both M^R and M^B on various example tiles. Red indicates high activation areas, while blue indicates low activation areas. M^B focuses more on closed glands (without lumen), while M^R focuses on open glands (with lumen).

and M^B (Figure 6). M^B and M^R focus primarily on closed and open glands, respectively. Similar to the results observed in Figure 5, this behavior suggests differences in glandular appearance within BXs and RPs and could be explained by mechanical alteration induced by the BX needle during tissue acquisition such as the cutting or deformation of glands during the needle puncture. The process of acquiring tissue could have other types of impact on the appearance of the corresponding images as well. While BXs are slivers of tissue traversing a shallow depth [3], RPs are planar sections potentially of the entire prostate [5]. Therefore, RPs contain tissue from a larger region of the prostate compared to BXs and can potentially contain a larger variety of tissue. Additionally, RPs are likely to contain tissue from different zones of the prostate, whereas BXs are less likely due to their smaller size. This difference in sample acquisition could also contribute to morphological differences between sample types. However, this is still a hypothesis and requires further investigation.

Quantitative evaluation confirmed some of the model performance patterns observed by expert pathologists. M^B had fewer FP and more FN in benign tiles on both RPs or BXs, while M^R had more FPs and fewer FNs (supplementary material, Figure S3). The larger number of FNs produced by M^B line up with observations from pathologists that M^B struggles to classify tiles with large cancerous glands. Similarly, the larger number of FPs produced by M^R line up with observations from pathologists that M^B struggles to classify tiles containing benign glands. This quantitative experiment also demonstrated that M^R had significantly better performance in both grades 3 and 4 cancers on RPs (supplementary material, Figure S3E,F) but did not have significantly different performance on BXs (supplementary material, Figure S3B,C). Again, this could indicate a difference in cancer appearance between the two sample types. Overall, this quantitative evaluation indicates that cancer grade had little effect on whether a tile was correctly classified, since both M^R and M^B performed similarly on grades 3 and 4 tiles in both BXs and RPs. This analysis also demonstrated that there were some differences between variant morphology within the RP and BX datasets. M^B was noted to perform worse on cribriform and ductal cancer tiles compared to M^R. This is most likely due to the fact that S^B contains fewer of these variant morphologies compared to S^R. Similarly, M^R performed worse on foamy cancer tiles (rare presentation of prostate cancer appearing similar to benign glands) compared to M^B, most likely because there are fewer tiles with this presentation in S^R compared to S^B (supplementary material, Table S2). However, given the small number of tiles with these presentations, variant morphology was likely not a driving factor in the performance differences observed in this study.

Model performance was further validated using a completed cooperative group phase III clinical trial from NRG Radiation Therapy Oncology Group (RTOG), NRG/RTOG 0521 [24]. This trial evaluated the added benefit of adjuvant docetaxel chemotherapy after androgen therapy and radiotherapy in a randomized controlled setting. The validation process on NRG/RTOG 0521 appeared to confirm the performance differences observed throughout the training process and sets. M^B was overwhelmingly preferred by expert pathologists compared to M^R on this set of BXs, with an even greater preference when the amount of cancer was deemed to be actionable for further analysis. Quantitative comparison of model performance on this external validation set confirmed these observations, with M^B performing significantly better than M^R across all measured metrics. Example outputs from both models (Figure 5E,F) demonstrated that M^B identified much more cancerous tissue compared to M^R. M^B does suffer from some FP errors on NRG/RTOG 0521 samples.

Training M^R and M^B using other architectures produced results similar to those observed with DenseNet, revealing that the observed performance differences on BXs and RPs were independent of model architecture. Additionally, results for M^{B+R} indicated that combining datasets during training did not result in better performance on either BXs or RPs compared to M^B and M^R and that sensitivity was particularly poor for BXs (Table 2).

Performance of all models on their own sample types aligned with similar findings reported in the literature. Pantanowitz [9] used an ensemble-based approach to predict slide-level classifications for prostate cancer BXs (sensitivity and specificity greater than 0.85). Tolkach presented similar models for RPs (F1 score of 0.94) [32]. Bulten [33] applied a similar method to achieve gland-level grade predictions on prostate cancer BXs, achieving a sensitivity and specificity greater than 0.88 for all cancer grades. Large challenges like the PANDAS challenge [11] have been developed for this very task, resulting in over 1,000 different cancer detection models for prostate BXs. Despite using a relatively simple neural network architecture for our models (M^B, M^R), they achieved a sensitivity and specificity similar to those reported in the aforementioned studies [9,11,32] when applied to the sample type they were trained with.

We acknowledge that our study has its limitations. First, S^B contains fewer samples and patients compared to S^R, which could have impacted model accuracy and generalizability. However, this design setup was necessary to ensure that both datasets originated from the same institution. Second, observations on gland size and appearance could not be quantified due to a lack of gland annotations. Future study will be dedicated to further characterizing these observations through quantitative measures of gland size and shape. Lastly, our study used 28 of the 750 slides within NRG/RTOG 0521 to conduct quantitative validation. This was on account of the challenge in acquiring manual annotations. However, qualitative evaluation of all models was performed on all 750 slide images. In future work we will seek to obtain additional ground truth annotations of the cancer extent on all remaining slide images, enabling a more extensive quantitative evaluation.

Despite the aforementioned limitations, this study reveals that sample type-specific AI models may be necessary for cancer detection in prostate histology slides. Models trained on a dataset of BXs should likely not be applied to RPs and vice versa due to distinct morphological differences between sample types. Substantial pretraining using self-supervised learning techniques [34,35] or other methods including foundation models [36] could prove useful for increasing model generalizability. However, given the core morphological differences between RPs and BXs identified in this study, it is unclear whether these approaches will be sufficient to improve generalizability. This must be validated in further studies.

Conclusions

Our results suggest that AI-based cancer detection models trained on BX samples are less than optimal when applied to RP specimens, and vice versa. A comprehensive review of model performance demonstrated that differential model performance appeared to be driven by morphological differences between BXs and RPs. These morphological differences could have originated from mechanical distortion of tissue architecture during BX acquisition. The resulting changes to tissue appearance could have been a driving factor in the model performance observed in this study. These findings suggest that cancer detection models trained on BXs do not readily generalize to RPs, and vice versa. This suggests a need to create sample-type specific cancer detection models for BXs and RPs. This is especially important as AI models are approved by the FDA and implemented in clinical settings [37]. Inappropriate invocation of these models could also have detrimental impact on other downstream tasks such as cancer grading [38], diagnosis [39], and prognosis [40]. In a clinical setting, this could negatively change clinical decision making and harm patients. At the very least, our findings suggest the need for further work to carefully evaluate sample-specific AI-based prostate cancer detection models.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute from grants UG1CA189867 (NCORP), U10CA180822 (NRG Oncology SDMC), U24CA196067 (NRG Specimen Bank), U10CA180868 (NRG Oncology Operations), and under award numbers R01CA268287A1, U01CA269181, R01CA26820701A1, R01CA249992-01A1, R01CA202752-01A1, R01CA208236-01A1, R01CA216579-01A1, R01CA220581-01A1, R01CA257612-01A1, 1U01CA239055-01, 1U01CA248226-01, 1U54CA254566-01, National Heart, Lung and Blood Institute 1R01HL15127701A1, R01HL15807101A1, National Institute of Biomedical Imaging and Bioengineering 1R43EB028736-01, VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service, Office of the Assistant Secretary of Defense for Health Affairs, through the DoD Breast Cancer Research Program Breakthrough Level 1 (W81XWH-19-1-0668), the Prostate Cancer Research Program (W81XWH-20-1-0851), the Lung Cancer Research Program (W81XWH-18-1-0440, W81XWH-20-1-0595), the Peer Reviewed Cancer Research Program (W81XWH-18-1-0404, W81XWH-21-1-0345, W81XWH-21-1-0160), the Kidney Precision Medicine Project (KPMP) Glue Grant, and sponsored research agreements from Bristol Myers-Squibb, Boehringer-Ingelheim, Eli-Lilly, and AstraZeneca. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the U.S. government.

Author contributions statement

BTF contributed to conceptualization, methodological development, software development, validation, formal data analysis, design of the investigation, resource allocation, data curation, writing in all stages, visualization of results and project administration. HMS and PL contributed to resource allocation and writing (reviewing and editing). MDF, MAH and RJ contributed to writing (reviewing and editing). JCS-R and RC allocated resources. TP contributed to validation, data curation, writing (review and editing) and project administration. TM contributed to formal data analysis, design of the investigation, data curation and writing (reviewing and editing). XF contributed to validation and writing (reviewing and editing). SC contributed to design of the investigation, resource allocation and supervision. AS worked on design of the investigation. JAE contributed to design of the investigation, resource allocation and writing (reviewing and editing). KH and MS worked on resource allocation and writing (reviewing and editing). SLP contributed to data curation, writing (reviewing and editing) and supervision. FF contributed to conceptualization, methodological development and supervision. AM contributed to conceptualization, methodological development, validation, formal data analysis, design of the investigation, resource allocation, writing (reviewing and editing), supervision, project administration and acquisition of funding. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work, which includes ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Data availability statement

Data used for this study are private and cannot be made available due to data use agreements between The University of Pennsylvania, Case Western, Emory Health, NRG/RTOG, and CorePlus. Code used for this study is available at the following GitHub repository: https://github.com/brennanFlannery/StressTestingDeepLearningModelsPCaDetection. Clinical trial registration for RTOG 0521: NCT00288080.

References

1. Siegel RL, Giaquinto AN, Ahmedin J, *et al*. Cancer statistics, 2024. *CA Cancer J Clin* 2024; **74**: 12–49.
2. American Cancer Society. Tests for Prostate Cancer|Prostate Cancer Diagnosis. [Accessed 20 February 2024]. Available from: https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/how-diagnosed.html
3. Noureldin ME, Connor MJ, Boxall N, *et al*. Current techniques of prostate biopsy: an update from past to present. *Transl Androl Urol* 2020; **9**: 1510–1517.
4. Munjal A, Leslie SW. Gleason score. In *Encyclopedia of Genetics, Genomics, Proteomics and Informatics.* StatPearls: Treasure Island, FL, 2022; 800.
5. Hugosson J, Stranne J, Carlsson SV. Radical retropubic prostatectomy: a review of outcomes and side-effects. *Acta Oncol* 2011; **50**: 92–97.
6. Humphrey PA. Histopathology of prostate cancer. *Cold Spring Harb Perspect Med* 2017; **7**: a030411.
7. Noguchi M, Stamey TA, McNeal JE, *et al*. Relationship between systematic biopsies and histological features of 222 radical prostatectomy specimens: lack of prediction of tumor significance for men with nonpalpable prostate cancer. *J Urol* 2001; **166**: 104–110.
8. da Silva LM, Pereira EM, Salles PGO, *et al*. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol* 2021; **254**: 147–158.
9. 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 Digit Health* 2020; **2**: e407–e416.

10. Sandbank J, Bataillon G, Nudelman A, *et al.* Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. *NPJ Breast Cancer* 2022; **8**: 129.
11. 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**: 154–163.
12. Leo P, Janowczyk A, Elliott R, *et al.* Computer extracted gland features from H&E predicts prostate cancer recurrence comparably to a genomic companion diagnostic test: a large multi-site study. *NPJ Precis Oncol* 2021; **5**: 35.
13. Bhargava HK, Leo P, Elliott R, *et al.* Computationally derived image signature of stromal morphology is prognostic of prostate cancer recurrence following prostatectomy in African American patients. *Clin Cancer Res* 2020; **26**: 1915–1923.
14. Leo P, Elliott R, Shih NNC, *et al.* Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: a multi-site study. *Sci Rep* 2018; **8**: 14918.
15. 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**: 2058–2066.
16. 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**: 1588–1595.
17. Eloy C, Marques A, Pinto J, *et al.* Artificial intelligence-assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. *Virchows Arch* 2023; **482**: 595–604.
18. 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**: 3383.
19. Hunter B, Hindocha S, Lee RW. The role of artificial intelligence in early cancer diagnosis. *Cancers (Basel)* 2022; **14**: 1524.
20. Janowczyk A, Zuo R, Gilmore H, *et al.* HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform* 2019; **3**: 1–7.
21. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019; **6**: 1–48.
22. Azevedo Tosta TA, de Faria PR, Neves LA, *et al.* Computational normalization of H&E-stained histological images: progress, challenges and future potential. *Artif Intell Med* 2019; **95**: 118–132.
23. Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks. *Proc IEEE Conf Comput Vis Pattern Recognit* 2017: 2261–2269.
24. Rosenthal SA, Hu C, Sartor O, *et al.* Effect of chemotherapy with docetaxel with androgen suppression and radiotherapy for localized high-risk prostate cancer: the randomized phase III NRG oncology RTOG 0521 trial. *J Clin Oncol* 2019; **37**: 1159–1168.
25. McInnes L, Healy J, Saul N, *et al.* UMAP: uniform manifold approximation and projection. *J Open Source Softw* 2018; **3**: 861.
26. Macenko M, Niethammer M, Marron JS, *et al.* A method for normalizing histology slides for quantitative analysis. *Proc IEEE Int Symp Biomed Imaging* 2009: 1107–1110.
27. Fernández-Moreno M, Lei B, Holm EA, *et al.* Exploring the trade-off between performance and annotation complexity in semantic segmentation. *Eng Appl Artif Intel* 2023; **123**: 106299.
28. Wang Q, Ma Y, Zhao K, *et al.* A comprehensive survey of loss functions in machine learning. *Ann Data Sci* 2022; **9**: 187–212.
29. Kingma DP, Ba JL. Adam: a method for stochastic optimization. In *The Thirteenth International Conference on Learning Representations.* ICLR: Appleton, WI, 2014.
30. Selvaraju RR, Cogswell M, Das A, *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020; **128**: 336–359.
31. Hyun JS. Clinical significance of prostatic calculi: a review. *World J Mens Health* 2018; **36**: 15–21.
32. Tolkach Y, Dohmgörgen T, Toma M, *et al.* High-accuracy prostate cancer pathology using deep learning. *Nat Mach Intell* 2020; **2**: 411–418.
33. 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**: 233–241.
34. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng* 2022; **6**: 1346–1352.
35. Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Comput Sci* 2022; **8**: e1045.
36. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. *Med Image Anal* 2024; **91**: 102996.
37. United States Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. [Accessed 20 February 2024]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
38. Montironi R, Cheng L, Cimadamore A, *et al.* Narrative review of prostate cancer grading systems: will the Gleason scores be replaced by the grade groups? *Transl Androl Urol* 2021; **10**: 1530–1540.
39. Descotes JL. Diagnosis of prostate cancer. *Asian J Urol* 2019; **6**: 129.
40. Abudoubari S, Bu K, Mei Y, *et al.* Prostate cancer epidemiology and prognostic factors in the United States. *Front Oncol* 2023; **13**: 1142976.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Example tiles and different random stain variations of the same tiles

Figure S2. UMAP embedding of BX and RP tiles

Figure S3. Statistical analysis of cancer grade appearance in SB,V and SR,V

Table S1. Tile level performance metrics for cancer detection models trained using different network architectures

Table S2. Observed instances of variant morphology in $S^{B,V}$ and $S^{R,V}$ by three expert pathologists