
FINDING HOLES: PATHOLOGIST LEVEL PERFORMANCE USING AI FOR CRIBRIFORM MORPHOLOGY DETECTION IN PROSTATE CANCER

**Kelvin Szolnoky¹, Anders Blilie^{2,3}, Nita Mulliqi¹, Toyonori Tsuzuki⁴, Hemamali Samaratunga⁵,
Matteo Titus¹, Xiaoyi Ji¹, Sol Erika Boman^{1,6}, Einar Gudlaugsson², Svein Reidar Kjosavik^{7,8},
José Asenjo⁹, Marcello Gambacorta¹⁰, Paolo Libretti¹⁰, Marcin Braun¹¹, Radzisław Kordek¹²,
Roman Łowicki¹², Brett Delahunt^{13,14}, Kenneth A. Iczkowski¹⁵, Theo van der Kwast¹⁶, Geert J.
L. H. van Leenders¹⁷, Katia R. M. Leite¹⁸, Chin-Chen Pan¹⁹, Emiel Adrianus Maria
Janssen^{2,20,21}, Martin Eklund¹, Lars Egevad¹⁴, and Kimmo Kartasalo²²**

¹Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

²Department of Pathology, Stavanger University Hospital, Stavanger, Norway

³Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

⁴Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagoya, Japan

⁵Aquesta Uropathology and University of Queensland, Brisbane, Queensland, Australia

⁶Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

⁷The General Practice and Care Coordination Research Group, Stavanger University Hospital,
Stavanger, Norway

⁸Department of Global Public Health and Primary Care, Faculty of Medicine, University of Bergen,
Bergen, Norway

⁹Department of Pathology, SYNLAB, Madrid, Spain

¹⁰Department of Pathology, SYNLAB, Brescia, Italy

¹¹Department of Pathology, Chair of Oncology, Medical University of Lodz, Lodz, Poland

¹²1st Department of Urology, Medical University of Lodz, Lodz, Poland

¹³Malaghan Institute of Medical Research, Wellington, New Zealand

¹⁴Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden

¹⁵Department of Pathology and Laboratory Medicine, University of California - Davis Health,
Sacramento, CA, USA

¹⁶Laboratory Medicine Program and Princess Margaret Cancer Center, University Health Network,
University of Toronto, Toronto, ON, Canada

¹⁷Department of Pathology, Erasmus MC, University Medical Center, Rotterdam, the Netherlands

¹⁸Department of Urology, Laboratory of Medical Research, University of São Paulo Medical School,
Sao Paulo, Brazil

¹⁹Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei,
Taiwan

²⁰Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger,
Stavanger, Norway

²¹Institute for Biomedicine and Glycomics, Griffith University, Brisbane, Queensland, Australia

²²Department of Medical Epidemiology and Biostatistics, SciLifeLab, Karolinska Institutet,
Stockholm, Sweden

October 13, 2025

ABSTRACT

Background: Cribriform morphology in prostate cancer is a histological feature that indicates poor prognosis and contraindicates active surveillance. However, it remains underreported and subject to significant interobserver variability amongst pathologists. We aimed to develop and validate an AI-based system to improve cribriform pattern detection.

Methods: We created a deep learning model using an EfficientNetV2-S encoder with multiple instance learning for end-to-end whole-slide classification. The model was trained on 640 digitised prostate core needle biopsies from 430 patients, collected across three cohorts. It was validated internally (261 slides from 171 patients) and externally (266 slides, 104 patients from three independent cohorts). Internal validation cohorts included laboratories or scanners from the development set, while external cohorts used completely independent instruments and laboratories. Annotations were provided by three expert uropathologists with known high concordance. Additionally, we conducted an inter-rater analysis and compared the model's performance against nine expert uropathologists on 88 slides from the internal validation cohort.

Results: The model showed strong internal validation performance (AUC: 0.97, 95% CI: 0.95-0.99; Cohen's kappa: 0.81, 95% CI: 0.72-0.89) and robust external validation (AUC: 0.90, 95% CI: 0.86-0.93; Cohen's kappa: 0.55, 95% CI: 0.45-0.64). In our inter-rater analysis, the model achieved the highest average agreement (Cohen's kappa: 0.66, 95% CI: 0.57-0.74), outperforming all nine pathologists whose Cohen's kappas ranged from 0.35 to 0.62.

Conclusion: Our AI model demonstrates pathologist-level performance for cribriform morphology detection in prostate cancer. This approach could enhance diagnostic reliability, standardise reporting, and improve treatment decisions for prostate cancer patients.

1 Introduction

Cribriform morphology in prostate cancer indicates increased metastatic potential, and is associated with adverse outcomes and increased mortality [1, 2]. The term cribriform comes from the Latin *cribrum*, meaning *sieve*, which describes its appearance where malignant epithelial cells form sheets punctured by sieve-like spaces [3, 4]. By definition, cribriform morphology is classified as at least Gleason pattern 4 [3]. In core needle biopsies, the prevalence of cribriform morphology ranges from 4% (for Gleason 3+4) up to 21% (for higher grade tumors) [5–7]. Given its prognostic value, the presence of cribriform morphology now contraindicates active surveillance strategies in prostate cancer management [8].

Despite this clinical importance, cribriform morphology remains underreported in routine practice [5]. This creates gaps in patient risk stratification. Furthermore, like Gleason grading, identifying cribriform patterns shows substantial interobserver variability and requires specialist expertise for consistent identification [7]. These diagnostic challenges are compounded by increasing workload pressures in pathology departments. Rising case volumes and declining number of specialists are stretching resources [9].

While AI solutions have emerged to address workload challenges, current approaches fail to fully meet the spectrum of diagnostic needs. Many AI models for prostate cancer focus solely on Gleason score [10]. However, comprehensive pathological reporting requires additional features beyond Gleason scoring. An effective AI solution must recognise and report multiple pathological features from a biopsy, with cribriform morphology detection being particularly important.

Currently, no study has sufficiently validated an AI-based system for cribriform detection. This study therefore aims to develop and validate an AI model for the automatic detection of cribriform morphology in prostate core needle biopsies.

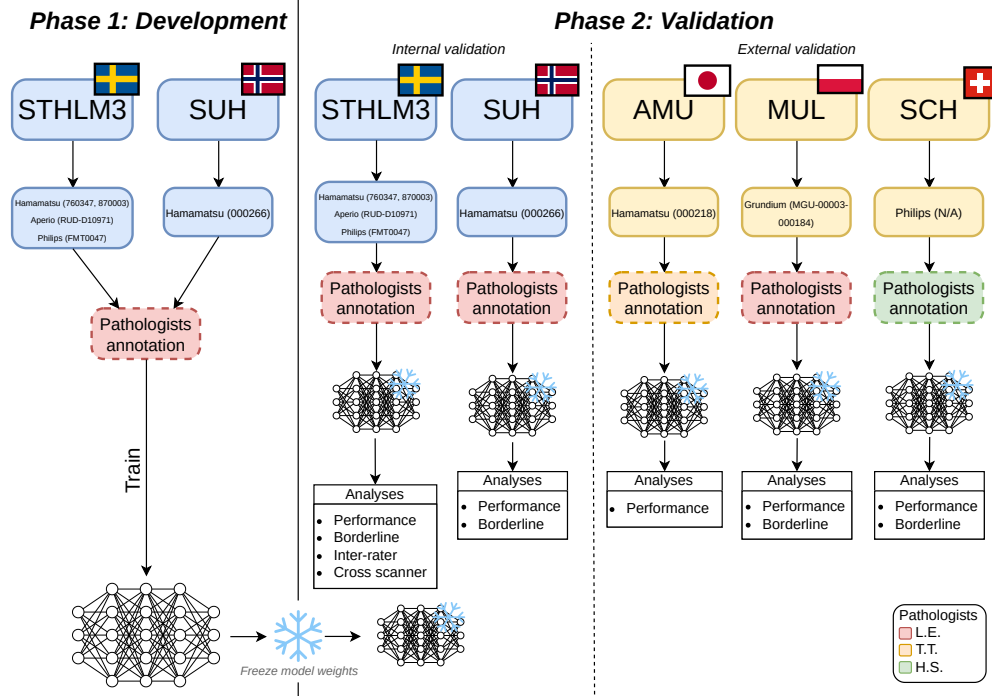


Figure 1. Overview of the study design. Phase 1 (Development) used subsets of the STHLM3 and SUH cohorts for model training. Phase 2 (Validation) included internal validation on reserved STHLM3/SUH data and external validation on three independent cohorts (AMU, MUL, SCH). Slides were digitised on scanners from multiple vendors and annotated by three pathologists. Numbers in parentheses indicate scanner serial numbers. Serial numbers for the scanners used at SCH are unavailable, but these scanners are distinct from those used in the other cohorts. No scanners in the external cohorts were present in the training data. Performance evaluation included standard metrics (AUC, Cohen’s kappa, sensitivity, specificity), inter-rater analysis comparing our model with nine pathologists, cross-scanner reproducibility assessment, and borderline case analysis. *Definition of abbreviations:* AUC = Area under the receiver operating characteristic curve.

2 Materials and Methods

We conducted a retrospective study in two phases: (1) model development and (2) validation (Figure 1). The study protocol has been published [11].

2.1 Data and Participants

We digitised formalin-fixed paraffin-embedded (FFPE), haematoxylin and eosin-stained prostate core needle biopsy slides from six cohorts: the Stockholm3 (STHLM3) trial [12], Capio S:t Görän Hospital, Sweden (STG), Stavanger University Hospital, Norway (SUH), Aichi Medical University, Japan (AMU), Medical University of Lodz, Poland (MUL), and Synlab Switzerland (SCH). Complete information about the cohorts – including collection dates, participants, and sampling methods – can be found in the protocol [11]. The slides were digitised using eight scanner instruments from four different vendors, including Philips (STHLM3, SCH), Grundium (MUL), Hamamatsu (AMU, STHLM3, SUH), and Aperio (STHLM3, STG). Some slides were scanned multiple times using different scanners. Please refer to Table 2 in the protocol for details regarding slide digitisation across cohorts [11].

Parts of the STHLM3, STG, and SUH cohorts were used for training and tuning the model during phase 1 (model development), while a portion of data from these cohorts was reserved for internal validation during phase 2 (model validation). The AMU, MUL, and SCH cohorts were used entirely

for external validation during phase 2. The validation datasets used during phase 2 were completely independent from the development process in phase 1 and only used for validation once. In other words, after phase 1, the model remained entirely fixed (frozen) throughout phase 2 without any adjustments. We defined validation cohorts as “internal” when their laboratory and/or specific scanner instrument had been included in the development set, while “external” cohorts contained samples from physical scanner instruments and laboratories that were completely independent from those used in development. All data partitions used to separate development from validation sets were grouped at the patient level to prevent data leakage, ensuring that slides from the same patient never appeared in both training and validation datasets.

2.2 Outcome

Cribriform morphology was defined per ISUP 2021 consensus as confluent malignant epithelial cells with multiple glandular lumina visible at low power (x10 objective), without intervening stroma or mucin between glandular structures [4]. Cribriform growth was annotated irrespective of whether it was invasive (within acinar adenocarcinoma) or non-invasive (intraductal carcinoma). This was justified by both forms often being assessed and reported together for prognostication and treatment planning, a practice supported by the 2019 ISUP consensus [13].

To minimise interobserver variability, we established a reference standard based on annotations from the lead pathologist (L.E.) or other experienced uropathologists (H.S., T.T.) whose concordance has been quantified in earlier studies [7]. To reduce the annotation burden for the reference standard pathologists, non-reference standard pathologists initially reviewed cases with Gleason pattern 4 to identify suspect slides with cribriform morphology. These preliminary annotations were used to upsample suspect cribriform cases for subsequent reference standard annotation. A non-reference standard pathologist was defined as one whose concordance to the lead pathologist (L.E.) is unknown. Non-reference standard annotations were not used to assess model performance.

For the STHLM3 and STG cohorts, a collection of 700 slides containing Gleason pattern 4 was assessed and annotated by the lead pathologist. In the other cohorts (SUH, MUL, and SCH), initial annotations were made by non-reference standard pathologists. A sample, with positive cases upsampled, was re-labelled by a reference standard pathologist. Detailed annotation protocols for each cohort are provided in Table A1 and the protocol [11].

For STHLM3 and STG, pixel-level annotations for glands representing cribriform morphology were made. For the other cohorts, only slide-level labels were annotated. When establishing the reference standard, L.E. (on STHLM3, SUH, and MUL) and H.S. (on SCH) also indicated cases they considered borderline cribriform. The term borderline was used for cases with features suggestive of cribriform growth that did not fully meet established morphological criteria. This category was intended to capture diagnostically difficult cases to permit statistical analyses on this specific substratum.

For a subset of the STHLM3 internal validation data, we also have annotations from the nine expert uropathologists included in an earlier interobserver reproducibility study [7].

2.3 Model Development

We extracted smaller images, referred to as patches, from each whole slide image (WSI) for input into the model. Each patch measured 256 by 256 pixels at 1 μm per pixel (10x magnification) and overlapped with neighbouring patches by 50% both vertically and horizontally. Patches with tissue covering less than 10% of the image were discarded based on tissue segmentation masks. An in-house segmentation model built on UNet++ with a ResNeXt-101 (32x4d) encoder was used to create the tissue segmentation masks [14].

We developed a multiple instance learning (MIL) model using an EfficientNetV2-S neural network backbone to detect cribriform morphology. The model processes WSIs by using the extracted patches, treating each slide as a bag of patches. These patches are processed through the neural network to extract patch-level features, which are aggregated via a gated attention mechanism to create slide-level features. To enhance generalisability, we implemented extensive data augmentation techniques. The final model utilised an ensemble of 10 models trained during 10-fold cross-validation, and used test-time augmentation for final predictions. Further details are available in the supplement (Section B).

2.4 Statistical Analysis

All analyses were prespecified [11]. We performed analyses at both individual cohort levels and aggregated internal and external cohort levels. Model performance was evaluated using receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC). An operating point of 0.5 was used for binary classification. We calculated sensitivity and specificity. We also measured the agreement between the model and the reference standard using Cohen’s kappa. For glass slides that were digitised multiple times in the STHLM3 cohort, performance metrics were calculated using only the original WSI that was annotated by the pathologist, rather than including all digital copies of the same physical slide. The 95% confidence intervals (CIs) for all metrics were calculated from nonparametric bootstrapping using 1,000 bootstrap samples. We created visual calibration plots to assess model calibration.

Using annotations on a subset of STHLM3 data from nine pathologists and our model, we conducted an inter-rater variability analysis to compare the model’s performance against expert pathologists. For each rater, including our model, we calculated the mean pairwise Cohen’s kappa coefficient against the other pathologists to quantify agreement levels. Furthermore, we conducted sensitivity analyses to evaluate cross-scanner reproducibility by calculating the pairwise Cohen’s kappa between model predictions on different digital scans of the same glass slides. This analysis used slides from the STHLM3 cohort that had been digitised multiple times using scanners from four vendors (Aperio, Grundium, Hamamatsu, Philips). Lastly, in an exploratory analysis, we quantified the prevalence of “borderline” cases in both true negative and false positive groups using annotations from L.E. and H.S. In analyses not specifically focused on borderline cases, these were classified as negative.

3 Results

3.1 Dataset characteristics

Patient and slide characteristics are summarised in Table 1 and A5. The study included a total of 705 patients: 430 in the training set, 171 in the internal validation set, and 104 in the external validation set (Table A4). Training was done on 1,280 WSIs from 640 physical slides. The internal validation cohorts included 211 physical slides from STHLM3 and 50 from SUH. The external validation cohorts contained 137 slides from MUL and 56 from SCH. The prevalence of cribriform pattern was higher in the external validation set (35%, $n = 94$) compared to training (24%, $n = 155$) and internal validation sets (24%, $n = 62$). The most common age interval was 65-69 years, comprising 40% of patients. Gleason score and ISUP grade distributions were relatively consistent across training, internal validation, and external validation sets.

3.2 Model Performance

On the internal validation set (STHLM3 and SUH cohorts), our deep learning model demonstrated an AUC of 0.97 (95% CI: 0.95, 0.99) and a Cohen’s kappa of 0.81 (95% CI: 0.72, 0.89), with a sensitivity of 0.92 (95% CI: 0.85, 0.98) and specificity of 0.93 (95% CI: 0.89, 0.96). For external validation (AMU, MUL, and SCH), the model achieved an AUC of 0.90 (95% CI: 0.86, 0.93) and a Cohen’s kappa of 0.55 (95% CI: 0.45, 0.64), with a sensitivity of 0.90 (95% CI: 0.84, 0.96) and specificity of 0.70 (95% CI: 0.63, 0.76). The ROC curves and confusion matrices illustrating these performance differences are presented in Figure A2 and 2a, while AUC, Cohen’s kappa, sensitivity, and specificity for all included cohorts are presented in Table 2.

Examining individual cohorts (Table 2), performance varied across datasets. STHLM3 achieved an AUC of 0.96 (95% CI: 0.94, 0.99) and a Cohen’s kappa of 0.80 (95% CI: 0.69, 0.90), while SUH demonstrated similar results with an AUC of 0.98 (95% CI: 0.95, 1.0) and a similar Cohen’s kappa of 0.80 (95% CI: 0.63, 0.96). Performance in external validation cohorts was more variable. The SCH cohort maintained results comparable to internal validation, with an AUC of 0.95 (95% CI: 0.87, 0.99) and a Cohen’s kappa of 0.71 (95% CI: 0.40, 0.93). However, while the AMU and MUL cohorts preserved good discriminative ability with AUCs of 0.92 (95% CI: 0.86, 0.97) and 0.89 (95% CI: 0.83, 0.94) respectively, their agreement metrics were notably lower, with Cohen’s kappa values of 0.42 (95% CI: 0.27, 0.60) for AMU and 0.53 (95% CI: 0.39, 0.65) for MUL.

Table 1. Patient and slide characteristics across all cohorts, showing demographic and clinical data.

Cohort	STG	STHLM3		SUH		AMU	MUL	SCH
Split	Train	Train	Test	Train	Test	Test	Test	Test
Patients								
<i>n</i>	67	287	140	76	31	43	49	12
Age, years								
≤49	0 (0%)	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)
50-54	1 (2%)	17 (6%)	6 (4%)	3 (4%)	1 (3%)	0 (0%)	1 (2%)	0 (0%)
55-59	2 (4%)	31 (11%)	18 (13%)	3 (4%)	2 (6%)	0 (0%)	1 (2%)	3 (25%)
60-64	4 (9%)	80 (28%)	35 (25%)	14 (18%)	1 (3%)	0 (0%)	6 (12%)	3 (25%)
65-69	4 (9%)	149 (52%)	72 (51%)	14 (18%)	6 (19%)	0 (0%)	9 (18%)	3 (25%)
≥70	36 (77%)	10 (3%)	8 (6%)	42 (55%)	21 (68%)	0 (0%)	31 (63%)	3 (25%)
Missing	20	0	0	0	0	43	0	0
PSA, ng/mL								
<3	3 (7%)	45 (16%)	16 (11%)	4 (5%)	0 (0%)	1 (2%)	0 (0%)	0 (0%)
3-<5	0 (0%)	88 (31%)	56 (40%)	5 (7%)	1 (3%)	1 (2%)	0 (0%)	1 (11%)
5-<10	7 (16%)	76 (26%)	39 (28%)	34 (45%)	13 (42%)	11 (26%)	0 (0%)	4 (44%)
≥10	35 (78%)	78 (27%)	29 (21%)	32 (43%)	17 (55%)	30 (70%)	0 (0%)	4 (44%)
Missing	22	0	0	1	0	0	49	3
Whole Slide Images								
<i>n</i>*	79	1,051	608	150	50	73	137	56
Physical slides	79	411	211	150	50	73	137	56
Cribriform	27 (34%)	81 (20%)	43 (20%)	47 (31%)	19 (38%)	28 (38%)	55 (40%)	11 (20%)
Gleason score								
3 + 3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (5%)
3 + 4	0 (0%)	61 (15%)	25 (12%)	68 (45%)	13 (26%)	0 (0%)	21 (15%)	18 (32%)
3 + 5	1 (1%)	11 (3%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (9%)
4 + 3	2 (3%)	131 (32%)	74 (35%)	37 (25%)	16 (32%)	0 (0%)	38 (28%)	19 (34%)
4 + 4	17 (22%)	158 (38%)	73 (35%)	25 (17%)	15 (30%)	0 (0%)	35 (26%)	5 (9%)
4 + 5	27 (34%)	39 (9%)	34 (16%)	18 (12%)	4 (8%)	0 (0%)	28 (20%)	6 (11%)
5 + 3	0 (0%)	1 (<1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
5 + 4	19 (24%)	6 (1%)	2 (<1%)	2 (1%)	2 (4%)	0 (0%)	15 (11%)	0 (0%)
5 + 5	13 (16%)	4 (<1%)	2 (<1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing	0	0	0	0	0	73	0	0
ISUP								
1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (5%)
2	0 (0%)	61 (15%)	25 (12%)	68 (45%)	13 (26%)	0 (0%)	21 (15%)	18 (32%)
3	2 (3%)	131 (32%)	74 (35%)	37 (25%)	16 (32%)	0 (0%)	38 (28%)	19 (34%)
4	18 (23%)	170 (41%)	74 (35%)	25 (17%)	15 (30%)	0 (0%)	35 (26%)	10 (18%)
5	59 (75%)	49 (12%)	38 (18%)	20 (13%)	6 (12%)	0 (0%)	43 (31%)	6 (11%)
Missing	0	0	0	0	0	73	0	0

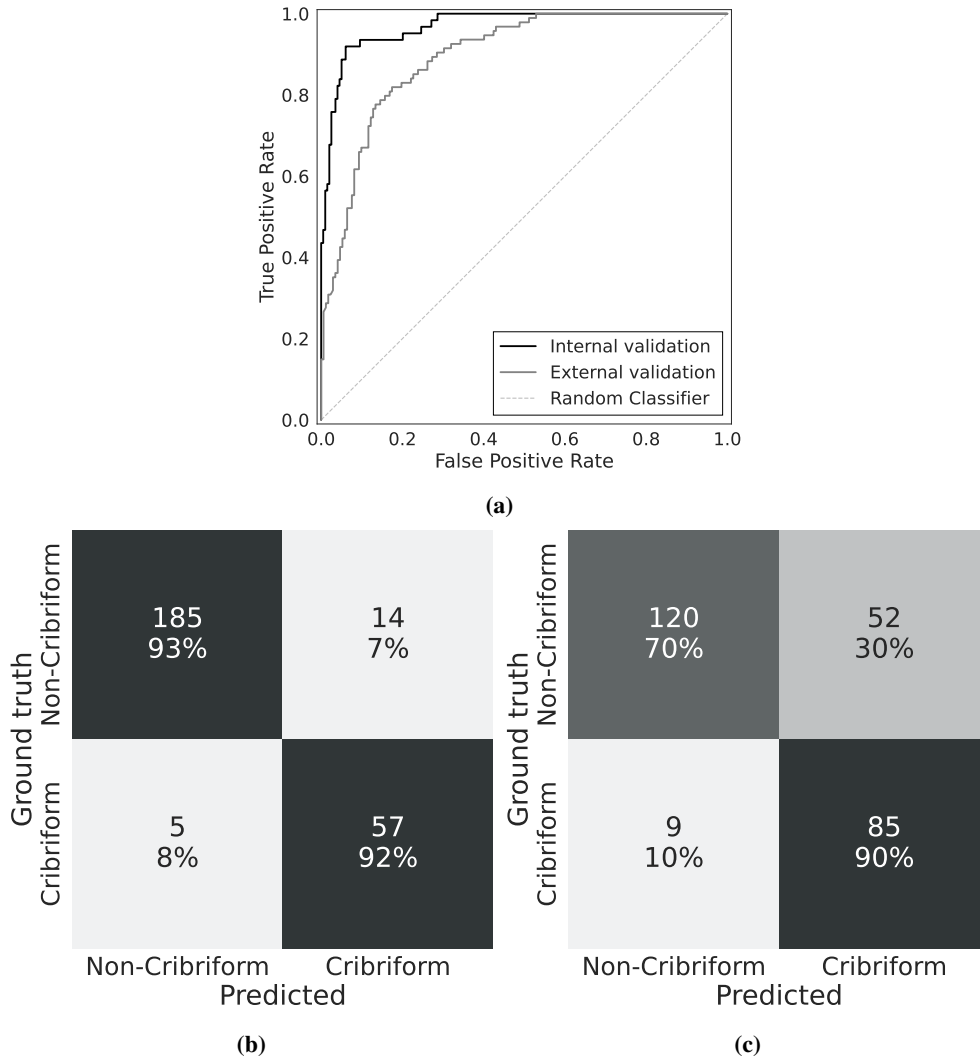
* Total number of whole slide images (digital copies of physical slides). This may exceed the number of physical slides when slides from a cohort were scanned multiple times on different scanners.

Definition of abbreviations: PSA = Prostate specific antigen; ISUP = International Society of Urological Pathology Grade.

Table 2. Performance metrics across all cohorts, including AUC, Cohen’s kappa, sensitivity, and specificity values with 95% confidence intervals. Type indicates the cohort’s validation status.

Cohort	Type	AUC	Cohen’s kappa	Sensitivity	Specificity
STHLM3	Internal	0.96 (0.94, 0.99)	0.8 (0.69, 0.9)	0.88 (0.78, 0.98)	0.95 (0.91, 0.98)
SUH	Internal	0.98 (0.95, 1.0)	0.8 (0.63, 0.96)	1.0 (1.0, 1.0)	0.84 (0.7, 0.97)
AMU	External	0.92 (0.86, 0.97)	0.42 (0.27, 0.6)	1.0 (1.0, 1.0)	0.49 (0.33, 0.63)
MUL	External	0.89 (0.83, 0.94)	0.53 (0.39, 0.65)	0.89 (0.8, 0.97)	0.67 (0.56, 0.77)
SCH	External	0.95 (0.87, 0.99)	0.71 (0.4, 0.93)	0.73 (0.43, 1.0)	0.96 (0.89, 1.0)
Overall	Internal	0.97 (0.95, 0.99)	0.81 (0.72, 0.89)	0.92 (0.85, 0.98)	0.93 (0.89, 0.96)
Overall	External	0.9 (0.86, 0.93)	0.55 (0.45, 0.64)	0.9 (0.84, 0.96)	0.7 (0.63, 0.76)

Definition of abbreviations: AUC = Area under the receiver operating characteristic curve.

**Figure 2.** (a) Receiver operating characteristic curves showing model performance on internal and external validation sets. (b) Confusion matrix on predictions for the internal validation set (STHLM3 and SUH). (c) Confusion matrix on predictions for the external validation set (AMU, MUL, and SCH).

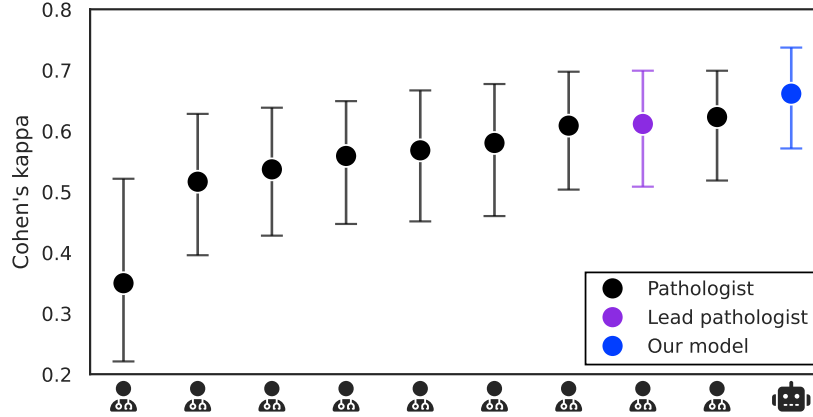


Figure 3. Pathologist concordance analysis comparing the agreement between our model (robot icon) and nine pathologists (physician icon), showing mean pairwise Cohen’s kappa values. For each rater, including our model, the mean pairwise Cohen’s kappa was calculated against the other pathologists only (the model was excluded from this average calculation). The whiskers indicate the 95% confidence interval. For exact values see Table A6.

The model showed good calibration internally, with predicted probabilities closely matching observed cribriform morphology (Figure A3a and A3b). In external validation, calibration deviated – especially at intermediate probabilities – leading to overdiagnosis and reduced specificity at the same operating point (0.5) used on the internal validation sets (Figure A1 and 2c).

In our exploratory analysis of borderline cases (Table A8), using annotations from L.E. (on STHLM3, SUH, and MUL) and H.S. (SCH), we found a significantly higher proportion of borderline cases among false positive cases (38%) compared to true negative cases (14%; Fisher’s exact test, $p = 0.008$). In the cross-scanner reproducibility analysis all scanners showed high concordance, with pairwise agreement ranging from 0.90 to 0.97. Scanner specific results for the cross-scanner reproducibility analysis are presented in the supplementary material (Section C).

3.3 Comparison with Pathologists

In the inter-rater analysis (Figure 3 and Table A6) we compared the model with nine pathologists on a subset of 88 slides from the STHLM3 validation cohort. Using the lead pathologist’s (L.E.) annotations as a reference, 43 slides were positive for the cribriform pattern. The model achieved the highest average pairwise Cohen’s kappa of 0.66 (95% CI: 0.57, 0.74). This exceeded the performance of all nine pathologists, whose average pairwise Cohen’s kappa values ranged from 0.35 (95% CI: 0.22, 0.52) to 0.62 (95% CI: 0.52, 0.70). The lead pathologist, who annotated the training data, ranked 3rd with an average pairwise Cohen’s kappa of 0.61 (95% CI: 0.51, 0.7).

4 Discussion

In this study, we developed and validated a deep-learning model to detect cribriform morphology in prostate cancer biopsies. Our model demonstrated strong discriminative performance across both internal and external cohorts, achieving high agreement with experienced pathologists’ annotations and, notably, the highest average agreement scores when compared against nine pathologists. However, in some external cohorts, a shift in calibration was observed, leading to an overdiagnosis of cribriform patterns. Even so, the model maintained acceptable performance levels comparable to those of a pathologist. Roughly 40% of the model’s false positive cases were considered to be borderline cribriform cases by the annotating pathologists. These findings suggest the model’s potential value as a screening tool for identifying high-risk regions within slides and helping to prioritise the most diagnostically challenging cases for expert pathologist review.

A strength of our study is the comprehensive validation strategy, which employed both internal and external validation cohorts alongside inter-rater and cross-scanner reproducibility analyses. However,

some limitations must be acknowledged. Specificity fell from 93% internally to 70% externally, indicating calibration and cohort shift. Inter-observer variability likely contributes, as prior work shows only moderate agreement among expert uropathologists assessing cribriform morphology (mean Cohen’s kappa ≈ 0.56) [7]. Additionally, both the cross-scanner reproducibility and pathologist concordance analyses were conducted on internal validation sets, potentially overestimating performance. These analyses were enabled by the availability of multi-rater annotations and repeatedly scanned slides from prior studies exclusively on the STHLM3 cohort.

To our knowledge, this is the first study to comprehensively validate an AI model specifically for cribriform pattern detection in prostate cancer. Previous research has primarily focused on Gleason grading or tumour detection [10]. Two earlier studies featured models for cribriform detection, but showed only modest results and lacked external multi-cohort validation [15, 16]. Our approach advances this work by developing a model that has been validated across multiple external, international cohorts and by comparing model performance directly against multiple pathologists.

The accurate detection of cribriform morphology represents one of the crucial decision points in treatment planning for prostate cancer patients. This prognostically significant pattern, though often overlooked in practice, directly influences risk stratification and treatment selection. Our model attempts to address this challenge by providing support for pathologists and enhancing diagnostic consistency. For pathologists confronting mounting caseloads, the model could offer assistance by highlighting regions of interest, streamlining workflows, and supporting diagnostic decisions. Improved reliability in cribriform detection could translate to better-informed treatment assessments for patients. Future research should focus on improving external calibration and conducting prospective clinical validation.

5 Conclusion

Our deep learning model demonstrates robust performance for automated cribriform morphology detection in prostate cancer, with performance comparable to experienced pathologists. This approach could enhance diagnostic reliability, standardise reporting of this prognostically important feature, and potentially improve treatment decisions for prostate cancer patients.

References

- [1] Kweldam CF, Wildhagen MF, Steyerberg EW, Bangma CH, Van Der Kwast TH, and Van Leenders GJ. Cribriform Growth Is Highly Predictive for Postoperative Metastasis and Disease-Specific Death in Gleason Score 7 Prostate Cancer. *Modern Pathology*. 2015;28(3):457–464. DOI: <https://doi.org/10.1038/modpathol.2014.116>
- [2] Russo GI, Soeterik T, Puche-Sanz I, Broggi G, Lo Giudice A, De Nunzio C, et al. Oncological Outcomes of Cribriform Histology Pattern in Prostate Cancer Patients: A Systematic Review and Meta-Analysis. *Prostate Cancer Prostatic Dis*. 2023;26(4):646–654. DOI: <https://doi.org/10.1038/s41391-022-00600-y>
- [3] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, and Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *American Journal of Surgical Pathology*. 2016;40(2):244–252. DOI: <https://doi.org/10.1097/PAS.0000000000000530>
- [4] Kwast TH van der, Leenders GJ van, Berney DM, Delahunt B, Evans AJ, Iczkowski KA, et al. ISUP Consensus Definition of Cribriform Pattern Prostate Cancer. *Am J Surg Pathol*. 2021;45(8):1118–1126. DOI: <https://doi.org/10.1097/PAS.0000000000001728>
- [5] Osiecki R, Kozikowski M, Białek Ł, Pyzlak M, and Dobruch J. The Presence of Cribriform Pattern in Prostate Biopsy and Radical Prostatectomy Is Associated with Negative Postoperative Pathological Features. *Cent European J Urol*. 2024;77(1):22–29. DOI: <https://doi.org/10.5173/ceju.2023.215>

- [6] Flammia S, Frisenda M, Maggi M, Magliocca FM, Ciardi A, Panebianco V, et al. Cribriform Pattern Does Not Have a Significant Impact in Gleason Score ≥ 7 /ISUP Grade ≥ 2 Prostate Cancers Submitted to Radical Prostatectomy. *Medicine*. 2020;99(38):e22156. DOI: <https://doi.org/10.1097/MD.00000000000022156>
- [7] Egevad L, Delahunt B, Iczkowski KA, Van Der Kwast T, Van Leenders GJLH, Leite KRM, et al. Interobserver Reproducibility of Cribriform Cancer in Prostate Needle Biopsies and Validation of International Society of Urological Pathology Criteria. *Histopathology*. 2023;82(6):837–845. DOI: <https://doi.org/10.1111/his.14867>
- [8] Ericson KJ, Wu SS, Lundy SD, Thomas LJ, Klein EA, and McKenney JK. Diagnostic Accuracy of Prostate Biopsy for Detecting Cribriform Gleason Pattern 4 Carcinoma and Intraductal Carcinoma in Paired Radical Prostatectomy Specimens: Implications for Active Surveillance. *Journal of Urology*. 2020;203(2):311–319. DOI: <https://doi.org/10.1097/JU.0000000000000526>
- [9] Märkl B, Füzesi L, Huss R, Bauer S, and Schaller T. Number of Pathologists in Germany: Comparison with European Countries, USA, and Canada. *Virchows Arch*. 2021;478(2):335–341. DOI: <https://doi.org/10.1007/s00428-020-02894-6>
- [10] Rabilloud N, Allaume P, Acosta O, De Crevoisier R, Bourgade R, Loussouarn D, et al. Deep Learning Methodologies Applied to Digital Pathology in Prostate Cancer: A Systematic Review. *Diagnostics*. 2023;13(16):2676. DOI: <https://doi.org/10.3390/diagnostics13162676>
- [11] Mulliqi N, Blilie A, Ji X, Szolnok K, Olsson H, Titus M, et al. Development and Retrospective Validation of an Artificial Intelligence System for Diagnostic Assessment of Prostate Biopsies: Study Protocol. *BMJ Open*. 2025;15(7):e097591. DOI: <https://doi.org/10.1136/bmjopen-2024-097591>
- [12] Grönberg H, Adolfsson J, Aly M, Nordström T, Wiklund P, Brandberg Y, et al. Prostate Cancer Screening in Men Aged 50–69 Years (STHLM3): A Prospective Population-Based Diagnostic Study. *The Lancet Oncology*. 2015;16(16):1667–1676. DOI: [https://doi.org/10.1016/S1470-2045\(15\)00361-7](https://doi.org/10.1016/S1470-2045(15)00361-7)
- [13] Leenders GJLH van, Kwast TH van der, Grignon DJ, Evans AJ, Kristiansen G, Kweldam CF, et al. The 2019 International Society of Urological Pathology (ISUP) Consensus Conference on Grading of Prostatic Carcinoma. *Am J Surg Pathol*. 2020;44(8):e87–e99. DOI: <https://doi.org/10.1097/PAS.0000000000001497>
- [14] Boman SE, Mulliqi N, Blilie A, Ji X, Szolnok K, Gudlaugsson E, et al. The Impact of Tissue Detection on Diagnostic Artificial Intelligence Algorithms in Digital Pathology. *arXiv [Preprint]* 2025. Available from: <https://arxiv.org/abs/2503.23021>
- [15] Ambrosini P, Hollemans E, Kweldam CF, Leenders GJLHV, Stallinga S, and Vos F. Automated Detection of Cribriform Growth Patterns in Prostate Histology Images. *Sci Rep*. 2020;10(1):14904. DOI: <https://doi.org/10.1038/s41598-020-71942-7>
- [16] Silva-Rodríguez J, Colomer A, Sales MA, Molina R, and Naranjo V. Going Deeper through the Gleason Scoring Scale: An Automatic End-to-End System for Histology Prostate Grading and Cribriform Pattern Detection. *Computer Methods and Programs in Biomedicine*. 2020;195:105637. DOI: <https://doi.org/10.1016/j.cmpb.2020.105637>

Ethical considerations

The study is conducted in agreement with the Declaration of Helsinki. The data were retrieved in one or more rounds at each of the participating international sites between 1 May 2012 and 1 May 2024. All data were deidentified at each site and provided to Karolinska Institutet in anonymised format. The centralised collection of patient samples from the international sites to Karolinska Institutet was approved by the Swedish Ethical Review Authority (permit 2019-05220). The following local approvals were provided to cover the data collection at each site: AMU (permit 2023-074 for the AMU cohort), Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32 for the STG and STHLM3 cohorts), the Bioethics Committee at the Medical University of Lodz (permit RNN/295/19/KE for the MUL cohort), and the Regional Committee for Medical and Health Research Ethics (REC) in Western Norway (permits REC/Vest 80924, REK 2017/71

for the SUH cohort). For the SCH cohort, ethical approval was waived by the respective local institutional review boards due to the retrospective usage of fully deidentified prostate specimens, and the data collection under the waiver was approved by the Swedish Ethical Review Authority (permit 2019-05220). Written informed consent was provided by the participants in the STHLM3 dataset.

Acknowledgements

The computations were made possible through the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at C3SE partially funded by the Swedish Research Council through grant agreement no. 2022-06725 and no. 2018-05973, and by the supercomputing resource Berzelius provided by the National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg Foundation.

We want to thank Carin Cavalli-Björkman for assistance with scanning and database support. We would also like to thank Silja Kavlie Fykse and Desmond Mfua Abono for scanning in Stavanger. We would like to acknowledge the patients who participated in the STHLM3 diagnostic study and the OncoWatch and NordCaP projects and contributed the clinical information that made this study possible.

A.B. received a grant from the Health Faculty at the University of Stavanger, Norway. M.E. received funding from the Swedish Research Council, Swedish Cancer Society, Swedish Prostate Cancer Society, Nordic Cancer Union, Karolinska Institutet, and Region Stockholm. K.K. received funding from the SciLifeLab & Wallenberg Data Driven Life Science Program (KAW 2024.0159), David and Astrid Hägelen Foundation, Instrumentarium Science Foundation, KAUTE Foundation, Karolinska Institute Research Foundation, Orion Research Foundation and Oskar Huttunen Foundation. L.E. received funding from the Swedish Cancer Foundation (23 2641 Pj) and the Stockholm Cancer Society (234053).

Author's Contributions

K.S., N.M., X.J., S.E.B. and K.K. developed the AI models. A.B., E.G., S.R.K., J.A., M.G., P.L., M.B., R.K., R.L., B.D., H.S., T.T., E.A.M.J., K.A.I., T.v.d.K., G.J.L.H.v.L., K.R.M.L., C.P., and L.E. collected, assessed and curated clinical datasets. N.M., X.J., K.S., S.E.B. and K.K. contributed to digitization, pre-processing and management of whole slide image data. K.S. conducted the statistical analyses. K.S. and K.K. analyzed and interpreted the study results. N.M. and K.K. acquired, optimized and maintained computing platforms. A.B., M.E. and K.K. acquired funding. L.E., M.E. and K.K. conceived of the study. K.K. takes responsibility for the integrity and accuracy of the analysis in this study. K.S. drafted the manuscript. All authors reviewed, edited and approved the manuscript.

SUPPLEMENTARY MATERIAL

FINDING HOLES: PATHOLOGIST LEVEL PERFORMANCE USING AI FOR CRIBRIFORM MORPHOLOGY DETECTION IN PROSTATE CANCER

Kelvin Szolnok¹, Anders Blilie^{2,3}, Nita Mulliqi¹, Toyonori Tsuzuki⁴, Hemamali Samaratunga⁵, Matteo Titus¹, Xiaoyi Ji¹, Sol Erika Boman^{1,6}, Einar Gudlaugsson², Svein Reidar Kjosavik^{7,8}, José Asenjo⁹, Marcello Gambacorta¹⁰, Paolo Libretti¹⁰, Marcin Braun¹¹, Radzisław Kordek¹², Roman Łowicki¹², Brett Delahunt^{13,14}, Kenneth A. Iczkowski¹⁵, Theo van der Kwast¹⁶, Geert J. L. H. van Leenders¹⁷, Katia R. M. Leite¹⁸, Chin-Chen Pan¹⁹, Emiel Adrianus Maria Janssen^{2,20,21}, Martin Eklund¹, Lars Egevad¹⁴, and Kimmo Kartasalo²²

¹Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

²Department of Pathology, Stavanger University Hospital, Stavanger, Norway

³Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

⁴Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagoya, Japan

⁵Aquesta Urothology and University of Queensland, Brisbane, Queensland, Australia

⁶Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

⁷The General Practice and Care Coordination Research Group, Stavanger University Hospital, Stavanger, Norway

⁸Department of Global Public Health and Primary Care, Faculty of Medicine, University of Bergen, Bergen, Norway

⁹Department of Pathology, SYNLAB, Madrid, Spain

¹⁰Department of Pathology, SYNLAB, Brescia, Italy

¹¹Department of Pathology, Chair of Oncology, Medical University of Lodz, Lodz, Poland

¹²1st Department of Urology, Medical University of Lodz, Lodz, Poland

¹³Malaghan Institute of Medical Research, Wellington, New Zealand

¹⁴Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden

¹⁵Department of Pathology and Laboratory Medicine, University of California - Davis Health, Sacramento, CA, USA

¹⁶Laboratory Medicine Program and Princess Margaret Cancer Center, University Health Network, University of Toronto, Toronto, ON, Canada

¹⁷Department of Pathology, Erasmus MC, University Medical Center, Rotterdam, the Netherlands

¹⁸Department of Urology, Laboratory of Medical Research, University of São Paulo Medical School, São Paulo, Brazil

¹⁹Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

²⁰Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway

²¹Institute for Biomedicine and Glycomics, Griffith University, Brisbane, Queensland, Australia

²²Department of Medical Epidemiology and Biostatistics, SciLifeLab, Karolinska Institutet, Stockholm, Sweden

October 13, 2025

A Figures and Tables

Table A1. Labelling and sampling methodology across different cohorts, sampling strategies, and the annotating pathologist (reference standard).

Cohort	Initial sampling*	Initial annotator	Second sampling*	Reference standard
STHLM3/STG	701 slides containing GP 4	L.E.	N/A	L.E.
SUH	332 slides containing GP 4	A.B.	120 ⁺ /40 ^{borderline} /40 ⁻ slides	L.E.
AMU	73 slides containing GP 4	T.T.	N/A	T.T.
MUL	276 slides containing GP 4	A.B.	74 ⁺ /63 ⁻ slides	L.E.
SCH	56 slides containing GP 4	Site pathologists	12 ⁺ /6 ⁻ blocks	H.S.

Definition of abbreviations: GP 4 = Gleason pattern 4.

* To enhance statistical power and reduce the annotation burden for the reference standards, a two-stage enrichment sampling strategy was employed. Initially, a non-reference standard pathologist annotated slides containing Gleason pattern 4. Subsequently, slides were resampled based on these preliminary annotations to enrich for potential cribriform patterns before final annotation by the reference standard pathologist. In the *second sampling* column, superscript symbols indicate the initial pathologist's assessment for cribriform and how sampling was done based off of these annotations. In the STHLM3, STG, and AMU cohorts the initial annotator was the reference standard, i.e. no second round of annotations was needed.

Table A2. Hyperparameters for the patch level model.

Hyperparameter	Value
Encoder	EfficientNetV2-S
Initial weights	Weights from Gleason scoring encoder
Loss function	Binary cross entropy loss (weighted)
Optimiser	AdamW
Learning rate	OneCycleLR scheduler (starting at $1 \cdot 10^{-5}$, peaking at $1 \cdot 10^{-4}$ after 1 epoch, and finally decreasing to $1 \cdot 10^{-6}$ following a cosine annealing schedule)
Weight decay	$1 \cdot 10^{-2}$
Batch size	64
Precision	bfloat16
Train augmentations	Random: crop, horizontal and vertical flip, 90 degrees rotation, colour jitter, gamma, tone curve, grey scale, unsharp mask or gaussian blur, ISO noise, gaussian noise, multiplicative noise, JPEG compression

Table A3. Hyperparameters for the slide level model.

Hyperparameter	Value
Encoder	EfficientNetV2-S
Initial weights	Cribriform patch level weights
Loss function	Binary cross entropy loss (weighted)
Optimiser	RAdam
Learning rate	$3 \cdot 10^{-5}$ (constant)
Weight decay	$1 \cdot 10^{-5}$
Batch size	1
Max bag size	2200
Precision	bfloat16
Train augmentations	Random: crop, horizontal and vertical flip, 90 degrees rotation, colour jitter, gamma, tone curve, grey scale, unsharp mask or gaussian blur, ISO noise, gaussian noise, multiplicative noise, JPEG compression
Test augmentations	Random: horizontal and vertical flip, 90 degrees rotation

Table A4. Patient and slide characteristics stratified by dataset split (training, internal validation, and external validation).

Split	Train	Internal test	External test	Overall
Patients				
<i>n</i>	430	171	104	705
Age, years				
≤49	0 (0%)	1 (<1%)	1 (2%)	2 (<1%)
50-54	21 (5%)	7 (4%)	1 (2%)	29 (5%)
55-59	36 (9%)	20 (12%)	4 (7%)	60 (9%)
60-64	98 (24%)	36 (21%)	9 (15%)	143 (22%)
65-69	167 (41%)	78 (46%)	12 (20%)	257 (40%)
≥70	88 (21%)	29 (17%)	34 (56%)	151 (24%)
Missing	20	0	43	63
PSA, ng/mL				
<3	52 (13%)	16 (9%)	1 (2%)	69 (11%)
3-<5	93 (23%)	57 (33%)	2 (4%)	152 (24%)
5-<10	117 (29%)	52 (30%)	15 (29%)	184 (29%)
≥10	145 (36%)	46 (27%)	34 (65%)	225 (36%)
Missing	23	0	52	75
Whole Slide Images				
<i>n</i>*	1,280	658	266	2,204
Physical slides	640	261	266	1,167
Cribriform	155 (24%)	62 (24%)	94 (35%)	311 (27%)
Gleason score				
3 + 3	0 (0%)	0 (0%)	3 (2%)	3 (<1%)
3 + 4	129 (20%)	38 (15%)	39 (20%)	206 (19%)
3 + 5	12 (2%)	1 (<1%)	5 (3%)	18 (2%)
4 + 3	170 (27%)	90 (34%)	57 (30%)	317 (29%)
4 + 4	200 (31%)	88 (34%)	40 (21%)	328 (30%)
4 + 5	84 (13%)	38 (15%)	34 (18%)	156 (14%)
5 + 3	1 (<1%)	0 (0%)	0 (0%)	1 (<1%)
5 + 4	27 (4%)	4 (2%)	15 (8%)	46 (4%)
5 + 5	17 (3%)	2 (<1%)	0 (0%)	19 (2%)
Missing	0	0	73	73
ISUP				
1	0 (0%)	0 (0%)	3 (2%)	3 (<1%)
2	129 (20%)	38 (15%)	39 (20%)	206 (19%)
3	170 (27%)	90 (34%)	57 (30%)	317 (29%)
4	213 (33%)	89 (34%)	45 (23%)	347 (32%)
5	128 (20%)	44 (17%)	49 (25%)	221 (20%)
Missing	0	0	73	73

* Total number of whole slide images (digital copies of physical slides). This may exceed the number of physical slides when slides from a cohort were scanned multiple times on different scanners.

Table A5. Patient and slide characteristics stratified by cribriform morphology status.

	Cribriform	Non-cribriform
Patients		
<i>n</i>	221	587
Age, years		
≤49	1 (<1%)	1 (<1%)
50-54	5 (3%)	27 (5%)
55-59	11 (6%)	54 (10%)
60-64	43 (22%)	123 (23%)
65-69	72 (37%)	219 (40%)
≥70	64 (33%)	120 (22%)
Missing	25	43
PSA, ng/mL		
<3	14 (8%)	60 (11%)
3-<5	26 (14%)	135 (26%)
5-<10	42 (23%)	165 (31%)
≥10	103 (56%)	166 (32%)
Missing	36	61
Whole Slide Images		
<i>n</i>*	544	1,660
Physical slides	311	856
Gleason score		
3 + 3	0 (0%)	3 (<1%)
3 + 4	17 (6%)	189 (23%)
3 + 5	3 (1%)	15 (2%)
4 + 3	86 (30%)	231 (28%)
4 + 4	112 (40%)	216 (27%)
4 + 5	53 (19%)	103 (13%)
5 + 3	0 (0%)	1 (<1%)
5 + 4	12 (4%)	34 (4%)
5 + 5	0 (0%)	19 (2%)
Missing	28	45
ISUP		
1	0 (0%)	3 (<1%)
2	17 (6%)	189 (23%)
3	86 (30%)	231 (28%)
4	115 (41%)	232 (29%)
5	65 (23%)	156 (19%)
Missing	28	45

* Total number of whole slide images (digital copies of physical slides). This may exceed the number of physical slides when slides from a cohort were scanned multiple times on different scanners.

Table A6. Mean pairwise Cohen's kappa values for our model and nine pathologists, evaluating 88 slides (43 annotated cribriform-positive by the lead pathologist) from the STHLM3 cohort. For each rater, including our model, the average was calculated against the pathologists only (the model was excluded from this average calculation). Values in parentheses indicate the 95% confidence interval.

Rater	Cohen's kappa
Our model	0.66 (0.57, 0.74)
Pathologist 1	0.62 (0.52, 0.7)
Lead pathologist	0.61 (0.51, 0.7)
Pathologist 3	0.61 (0.5, 0.7)
Pathologist 4	0.58 (0.46, 0.68)
Pathologist 5	0.57 (0.45, 0.67)
Pathologist 6	0.56 (0.45, 0.65)
Pathologist 7	0.54 (0.43, 0.64)
Pathologist 8	0.52 (0.4, 0.63)
Pathologist 9	0.35 (0.22, 0.52)

Table A7. Cross-scanner reproducibility analysis showing pairwise Cohen's kappa values between different scanner types for 71 slides (19 annotated cribriform-positive by the lead pathologist) from the STHLM3 validation set that were scanned on 4 different scanners. Values in parentheses indicate the 95% confidence interval.

Scanner	Aperio	Grundium	Hamamatsu	Philips	Average
Aperio	-	0.97 (0.82, 1.00)	0.97 (0.83, 1.00)	0.93 (0.79, 1.00)	0.96 (0.87, 0.99)
Grundium	0.97 (0.82, 1.00)	-	0.93 (0.77, 1.00)	0.97 (0.84, 1.00)	0.95 (0.87, 0.99)
Hamamatsu	0.97 (0.83, 1.00)	0.93 (0.77, 1.00)	-	0.90 (0.74, 0.97)	0.93 (0.81, 0.99)
Philips	0.93 (0.79, 1.00)	0.97 (0.84, 1.00)	0.90 (0.74, 0.97)	-	0.93 (0.80, 0.99)

Table A8. Analysis of borderline cases comparing the prevalence of borderline cribriform morphology, as annotated by two experienced uropathologists, between true negative and false positive predictions. Type indicates the cohort's validation status.

Cohort	Type	True Negatives			False Positives			p-value*
		n	Borderline	%	n	Borderline	%	
STHLM3	Internal	452	22	5%	24	10	42%	<0.001
SUH	Internal	26	1	4%	5	2	40%	0.06
MUL	External	55	7	13%	27	9	33%	0.038
SCH	External	43	7	16%	2	2	100%	0.036
Overall	Internal	478	23	5%	29	12	41%	<0.001
Overall	External	98	14	14%	29	11	38%	0.008

* Fisher's exact test

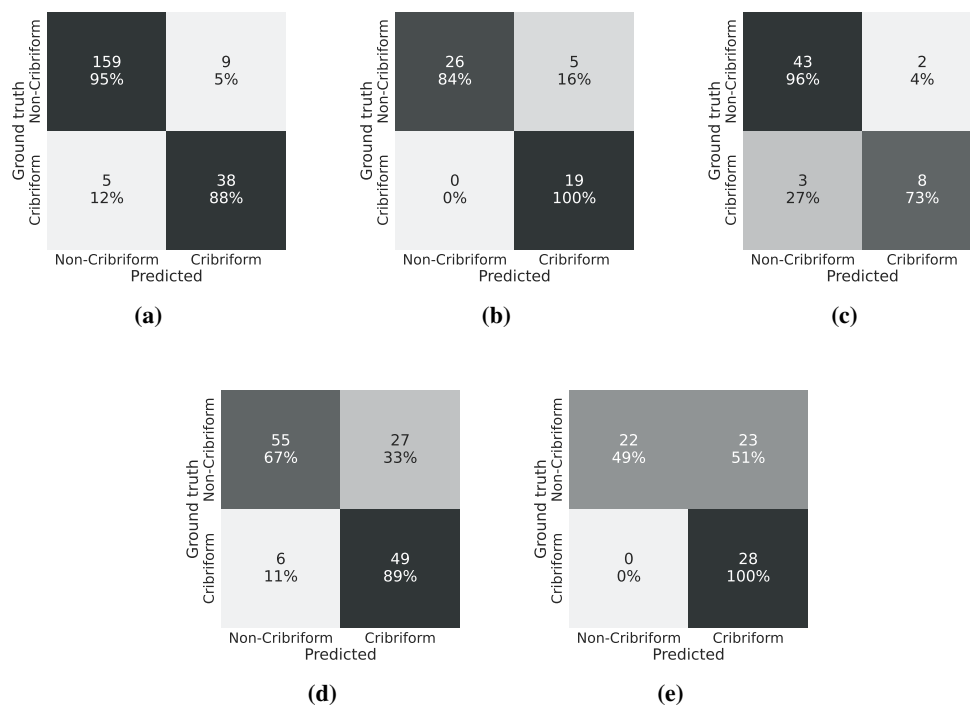


Figure A1. Confusion matrices on predictions for cohorts (a) STHLM3 (b) SUH (c) SCH (d) MUL (e) AMU

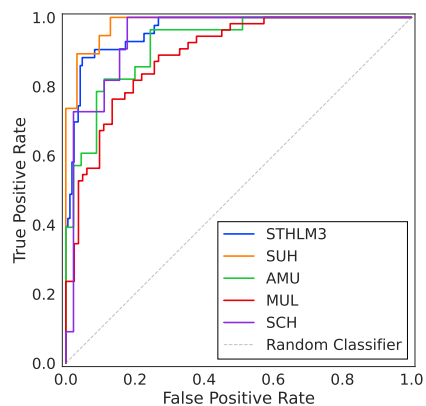


Figure A2. Receiver operating characteristic curves showing model performance for the different cohorts.

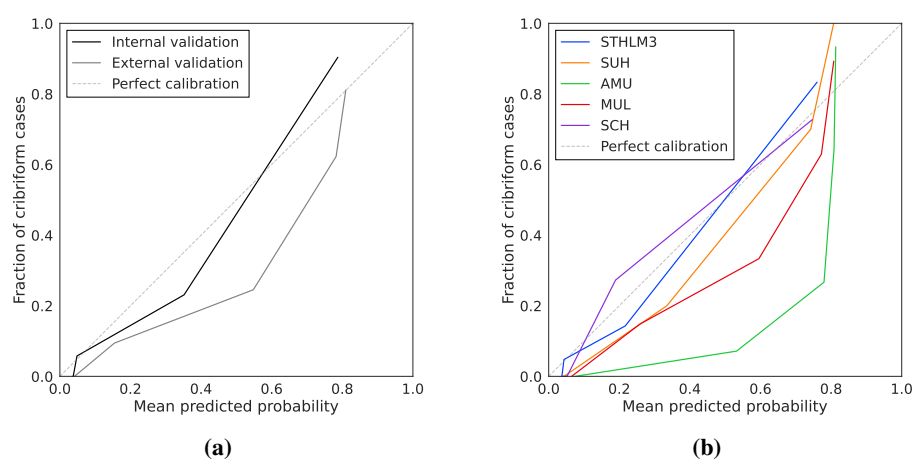


Figure A3. (a) Calibration curves demonstrating the relationship between predicted probabilities and observed frequencies of cribriform morphology. (b) Calibration curves demonstrating the relationship between predicted probabilities and observed frequencies of cribriform morphology for the different cohorts.

B Materials and Methods

B.1 Data Preparation

For STHLM3 and STG, pixel-wise annotations were made. The lead pathologist created pixel-wise annotations (marking cribriform regions) on only one digital version of each slide. Due to differences in how each scanner positioned the slide during digitisation, these annotations could not be directly transferred to other digital versions of the same slide. To address this limitation and increase our training data, we designed a simple phase correlation-based image registration algorithm. For slides with multiple scans, we created binary tissue segmentation masks and used a Fast Fourier Transform-based cross-correlation algorithm to align annotations across different digital versions of the same physical glass slide.

B.2 Model Development

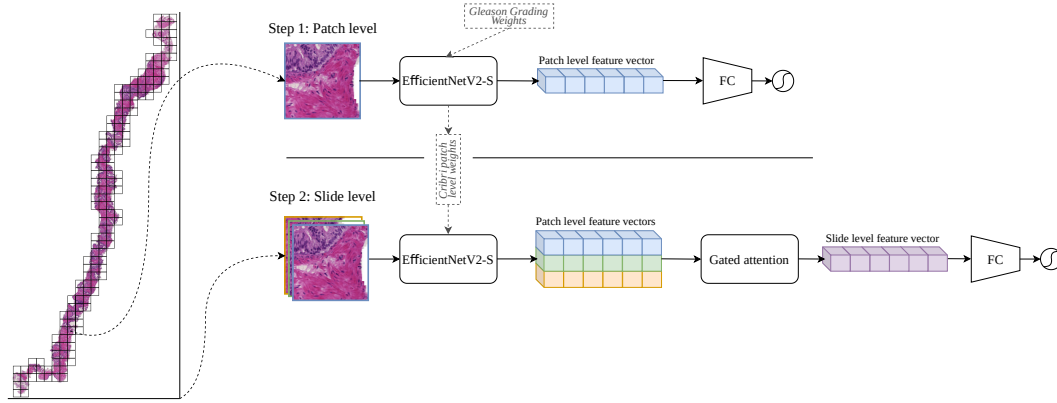


Figure A4. Model architecture illustrating the patch level and slide level classifiers for cribriform morphology detection in prostate cancer.

Definition of abbreviations: FC = Fully connected layer.

We implemented a two-step transfer learning procedure to enhance performance, convergence speed, and generalisation. The model architecture and transfer learning process are illustrated in Figure A4. This approach furthermore enabled the incorporation of SUH cohort data during Step 2, as this dataset lacked pixel-level annotations required for fully supervised training.

Step 1: Fully Supervised Patch Level Classifier. We created a patch level classifier using a convolutional neural network. EfficientNetV2-S was chosen as the backbone due to its high performance and relatively low resource usage [1]. To detect cribriform morphology at the patch level, we fine-tuned an EfficientNetV2-S encoder, which was previously part of a multiple instance learning (MIL) model trained on a large Gleason grading dataset [2]. The feature vector was passed through a fully connected layer with a sigmoid activation function, producing a probability score for cribriform morphology.

Step 2: Weakly Supervised Slide Level Classifier. We then developed a slide level classifier by transferring the cribriform patch level encoder weights into a MIL architecture. The encoder weights were not frozen, allowing for further fine-tuning during slide-level training. Bags of patches from a slide were processed through the encoder to create patch level feature vectors. These vectors were pooled together using gated attention to form a single slide level feature vector, which was then passed through a sequence of fully connected layers with normalisation, activation, and dropout, followed by a sigmoid activation function to produce a slide level probability score.

B.3 Training

We used a binary cross-entropy loss function, weighted by the frequency of positive labels in the training data, with a static probability threshold of 0.5 for classification. The AdamW optimiser was employed with a one cycle learning rate scheduler for the patch level model (starting at $1 \cdot 10^{-5}$,

peaking at $1 \cdot 10^{-4}$ after 1 epoch, and finally decreasing to $1 \cdot 10^{-6}$) and a constant learning rate of $3 \cdot 10^{-5}$ for the slide level model. The patch level classifier uses a weight decay of $1 \cdot 10^{-2}$ while the slide level model uses $1 \cdot 10^{-5}$. Data augmentations included random cropping, vertical and horizontal flips, colour and brightness jitter, sharpening, blurring, noise, JPEG compression, and random greyscale conversion. Furthermore, during training, for slides scanned multiple times on different scanners, we randomised which digital scan of a biopsy slide to use on each epoch. Hyperparameters are summarised in Tables A2 and A3. The models were trained for 8 and 32 epochs for the patch level and slide level respectively, with checkpoints every epoch, retaining only the checkpoint with the highest non-weighted Cohen's kappa on the hold-out fold. We used 10-fold cross-validation to evaluate the model during development. To avoid data leakage when transferring the Gleason grading weights, we employed protocol-defined splits. Platt scaling was applied by fitting a logistic regression model to the held-out folds in the training cross-validation.

For the patch level classifier, patches were labelled as cribriform-positive if they contained more than 2% of positively annotated tissue. Due to the constraints of needing pixel-wise annotations for the patch level classifier, the classifier could only be trained on data from the STHLM3 and STG cohorts. For the slide level classifier, bags of patches were labelled based on slide level annotations, enabling the classifier to be trained on the SUH cohort as well. To further utilise pixel-wise annotations in the STHLM3 and STG data, bags of patches from STHLM3 and STG were labelled as positive only if they contained a patch with cribriform morphology. For further regularisation, with the 50% overlap that the patches were extracted with, we could construct two sets of non-overlapping patches per WSI. For each forward pass of the model, one of these two non-overlapping sets was used. During validation, we used all extracted patches from a slide.

B.4 Inference

Our final model utilised a 10-fold ensemble approach derived from the cross-validation folds during model development. For inference, we extracted patches of size 256 by 256 pixels at a resolution of 1 μm per pixel, with 50% overlap between adjacent patches both vertically and horizontally. Patches were excluded if tissue content comprised less than 10% of the image. All extracted patches were passed through the model. To enhance prediction robustness, we applied test time augmentation with 5 iterations per ensemble model, using non-destructive transformations at a patch level, such as flipping and rotation. The final prediction was generated through soft voting, averaging predictions across all test time augmentation iterations and ensemble models.

B.5 Software and Hardware

Models and statistical analyses were implemented in Python 3.10 using Pytorch version 2.4 and Pytorch Lightning version 2.3. Support packages included Albumentations (1.4.12), LMDB (1.5.1), Numpy (2.1.3), Polars (1.14.0), Scipy (1.14.1), Scikit-learn (1.5.2), Timm (1.0.11). Plots were made using Matplotlib (3.9.2) and Seaborn (0.13.2). Models were trained using a single NVIDIA A100 80gb Tensor Core GPU. Inference was run on a single NVIDIA A100 40gb Tensor Core GPU. The extracted patches were encoded as JPEGs and saved to the Lightning Memory-Mapped Database (LMDB) format, as this allowed for efficient random reads which were needed in both phases of training. The resulting patch level and slide level models had 20.2 million and 21.9 million trainable parameters, respectively. The final training took 10 hours per data fold for a total of 100 GPU hours.

C Results

C.1 Cross-scanner Reproducibility

When examining cross-scanner reproducibility (Table A7), we utilised 71 slides from the STHLM3 internal validation set that had all been scanned on scanners from 4 different vendors. The subset included 19 slides that contained the cribriform pattern. The average pairwise Cohen's kappa values across scanners demonstrated high consistency, with Aperio achieving the highest average agreement at 0.96 (95% CI: 0.87, 0.99), followed by Grundium at 0.95 (95% CI: 0.87, 0.99), while Hamamatsu and Philips showed average Cohen's kappa values of 0.93 (95% CI: 0.78, 0.99) and 0.93 (95% CI: 0.81, 0.99) respectively. The highest level of agreement was observed between Aperio and Grundium scanners (Cohen's kappa: 0.97, 95% CI: 0.80, 1.00) and between Aperio and Hamamatsu scanners

(Cohen's kappa: 0.97, 95% CI: 0.82, 1.00). The lowest agreement was found between Hamamatsu and Philips scanners with a Cohen's kappa value of 0.90 (95% CI: 0.73, 0.97).

References

- [1] Tan M and Le QV. EfficientNetV2: Smaller Models and Faster Training. arXiv [Preprint] 2021. Available from: <http://arxiv.org/abs/2104.00298>
- [2] Mulliqi N, Blilie A, Ji X, Szolnoky K, Olsson H, Boman SE, et al. Foundation Models – A Panacea for Artificial Intelligence in Pathology? arXiv [Preprint] 2025. Available from: <http://arxiv.org/abs/2502.21264>