BMC Medical Imaging

RESEARCH Open Access

Unveiling key pathomic features for automated diagnosis and Gleason grade estimation in prostate cancer

Valentina Brancato[1*], Mario Verdicchio[1,2], Carlo Cavaliere[1], Francesco Isgrò[2], Marco Salvatore[1] and Marco Aiello[1]

Abstract

Background Recent advances in histology scanning technology and Artificial Intelligence (AI) offer great opportunities to support cancer diagnosis. The inability to interpret the extracted features and model predictions is one of the major issues limiting the acceptance of AI models in clinical practice, and a clear representation of the relevance of the extracted features and model predictions is lacking. Focusing on the problem of prostate cancer (PCa) diagnosis and grading, this study aims to detect which are the most discriminant features for distinguishing malignant from non-malignant tissue and Gleason patterns, leaving the evaluation of models' classification performances as a secondary goal.

Methods Utilizing a dataset of 187 annotated H&E-stained whole-slide images, the study explores three magnification levels, extracting 1971 features per tile. Two machine-learning classification tasks are conducted (Malignant vs. Non-malignant and High-grade vs. Low-grade) for each magnification. SHapley Additive exPlanations method was used to investigate models' interpretability, estimating the importance of pathomic features and their impact on prediction models.

Results Wavelet features were consistently prominent in "High-grade vs Low-grade" classification task, together with local binary pattern descriptors. Some histogram features appeared as key features for diagnostic classification tasks. The identified key discriminant features were classified for their specificity with respect to WSI magnification. Very high AUC values were reached for PCa diagnosis task ($0.97 < AUC < 0.99$), while task involving Low- versus High-grade classification exhibit lower AUC values (maximum AUC: 0.72–0.73).

Conclusions This work provides new insights for the explanation of hand-crafted pathomic-based classification models for PCa diagnosis and grading.

Clinical trial number Not applicable.

Keywords Prostate cancer, Gleason grade, Pathomics, Digital pathology, Decision support systems

*Correspondence:
Valentina Brancato
valentina.brancato@synlab.it
[1]IRCCS SYNLAB SDN, Via E. Gianturco 113, 80143 Naples, Italy
[2]Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Claudio 21, 80125 Naples, Italy

Introduction

Artificial intelligence (AI) is rapidly transforming the field of medical pathology, particularly in oncology, where the computational analysis of digital pathology images (pathomics) is garnering substantial interest for a wide range of applications. Pathomics converts digitized whole-slide images (WSIs) into analysable datasets using AI, linking pathological features to clinically related indicators. This innovative approach can improve tasks such as cancer classification, offering benefits like improved workflow, efficiency, reproducibility, cost reduction, and better training opportunities [1].

In pathomics, Machine Learning (ML) models are trained with datasets of WSIs that have been annotated and/or classified by pathologists [1].

ML-based techniques use predefined hand-crafted features from WSI patches or regions of interest to develop pattern-recognition criteria for classification tasks.

Deep learning (DL), an advanced subset of ML, has gained popularity for its ability to bypass hand-crafted features by using neural networks with multiple hidden layers, combining feature learning and classification. With sufficient data and techniques, DL identifies discriminative patterns, achieving high classification performance. However, limited labeled training data remains a challenge in the medical field.

Focusing on suitable clinical applications, one of the most important open challenges in the field of digital pathology is the accurate identification and grading of prostate cancer (PCa) on WSI [2].

This because PCa, other than being one of the most common and high-mortality cancers worldwide [3], is also characterized by an extremely high heterogeneity, both within and between individuals [4]. The Gleason grade is a key histological parameter for assessing PCa aggressiveness based on glandular structures in biopsy or prostatectomy specimens. It stratifies patients into five prognostic grade groups, determined by the two most predominant Gleason patterns and their corresponding score [5–7].

Figure 1 displays examples of tumor regions with different Gleason patterns and benign prostatic tissues,

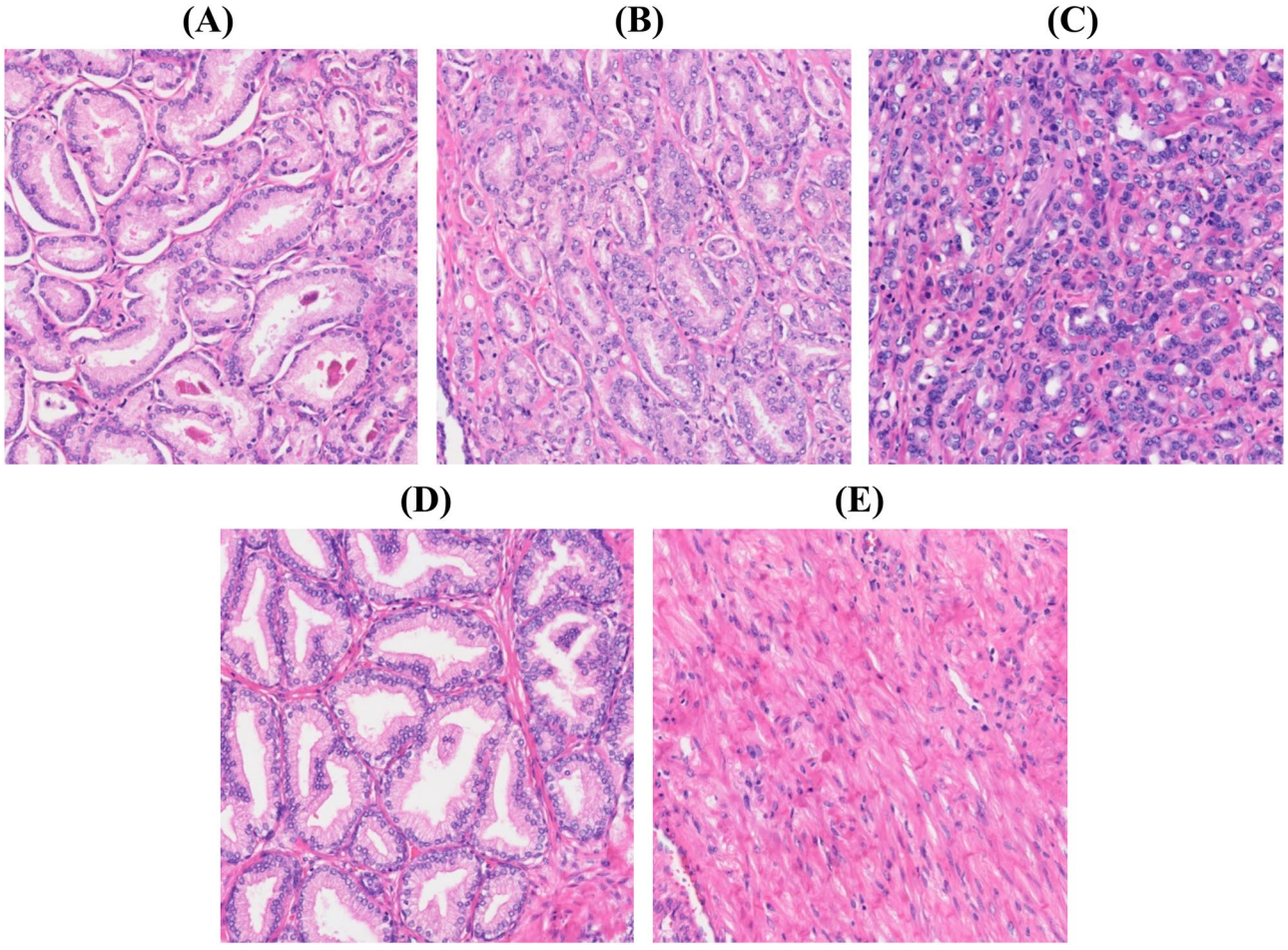

Fig. 1 Examples of prostatic cancerous and benign tissue regions. (**A**) well-formed glands (Gleason pattern 3); (**B**) poorly formed glands (Gleason pattern 4); (**C**) poorly formed glands and single cells (Gleason pattern 5); (**D**) Normal tissue; (**E**) Stroma

illustrating the textural variation between grades due to abnormal changes in gland structures.

Accurate Gleason grading is crucial for determining appropriate treatment but is time-consuming, labor-intensive, and prone to variability due to PCa's biological heterogeneity, posing risks of misdiagnosis and over-diagnosis [8].

Automating Gleason grading could significantly improve clinical decision-making, efficiency, and consistency.

Recent advances in histology scanning technology, AI and computer vision offer great opportunities to improve the PCa diagnosis and Gleason grading fidelity.

Several published studies on digital pathology have addressed automatic PCa diagnosis and grading [9, 10], with early approaches focusing on gland-nuclei-based methods leveraging glandular structures and nuclei characteristics within tissue samples [11]. However, these methods encountered notable challenges when confronted with the task of effectively classifying regions of high-grade carcinoma where some of the tissue structures (e.g. glands) cannot be differentiated.

In contrast, texture-based techniques operate at the tissue level, simplifying the computational process by avoiding intricate cellular or glandular segmentations. These methods translate textural patterns within WSIs into meaningful insights, directly impacting decision-making [12].

Recent advancements in deep learning (DL) have revolutionized PCa computational pathology, showing remarkable ability to identify complex patterns in pathology images [13]. However, DL models face challenges due to their lack of interpretability, often seen as "black boxes", hindering understanding of their classifications. Additionally, the need for large annotated datasets limits their applicability in settings with limited data [14].

Subsequent breakthroughs in the applicability of DL models toward reproducible PCa AI diagnostic tools started in 2019 with models trained and tested on very large cohorts [15–17].

Despite progress, further advancements are needed before widespread clinical adoption. A comprehensive solution requires understanding how these methods work together, leveraging their strengths while addressing their limitations. Of note, interpretability remains a key barrier to AI adoption in medicine.

Although many pathomic models have been developed and achieved good results, pathologists are often hesitant to adopt such models due to the unclear internal mechanisms or unexplained features [18].

This study focuses on well-established problems such PCa diagnosis and grading, which are areas where the scientific community is actively engaged.

A multiscale, multi-channel pathomic-based classification tasks was explored for PCa diagnosis and Gleason Grade classification from H&E-stained digital pathology images. The study design encompasses not only a thorough evaluation of the efficacy of the developed models in distinguishing malignant from non-malignant tissue and classifying different Gleason patterns, but also to detect which are the most discriminant features for automated diagnosis and Gleason grade estimation.

By delving into an extensive range of features, the aim is to provide novel pathomic insights into PCa diagnosis and grading.

Materials and methods

In this section, data used in this study, as well as the preprocessing, feature extraction and selection methods are introduced. Moreover, the classifier modelling and metrics used for the evaluation are provided. The overall methodologic workflow is shown in Fig. 2.

The minimum information about clinical AI modeling (MI-CLAIM) was used to provide transparency in the documentation of the developed algorithms, including an evaluation of bias and instructions for external reproducibility [19]. The MI-CLAIM checklist is provided in Supplementary Materials.

Dataset

A public dataset of H&E-stained whole slide image dataset of prostatectomy specimens was used (https://aggc22.grand-challenge.org/AGGC22/). It consists of 187 H&E-stained WSIs dataset of prostatectomy with annotations performed by experienced pathologists. The WSIs were acquired with brightfield imaging at 0.5 μm × 0.5 μm per pixel resolution (20× magnification).

Class labels of the annotations include Gleason pattern 3 (G3), Gleason pattern 4 (G4), Gleason pattern 5 (G5), normal prostatic tissue, and stroma tissue [10].

Preprocessing

To minimize the impact of staining variations on the computation of textural features, the Macenko normalization method, which utilizes color deconvolution, was employed [20]. The method was implemented in MATLAB using the following parameters: alpha = 1.0 (Optical Density -OD- pseudo-min/max tolerance), beta = 0.15 (OD threshold for transparent pixels), and Io = 255 (transmitted light intensity). A representative H&E-stained slide from the dataset was selected as the reference (target) image. Stain concentration vectors were normalized using the 99th percentile, and color reconstruction was performed via exponential transformation of the stain concentration and target stain matrix. This ensured consistent color representation across all patches prior to feature extraction.

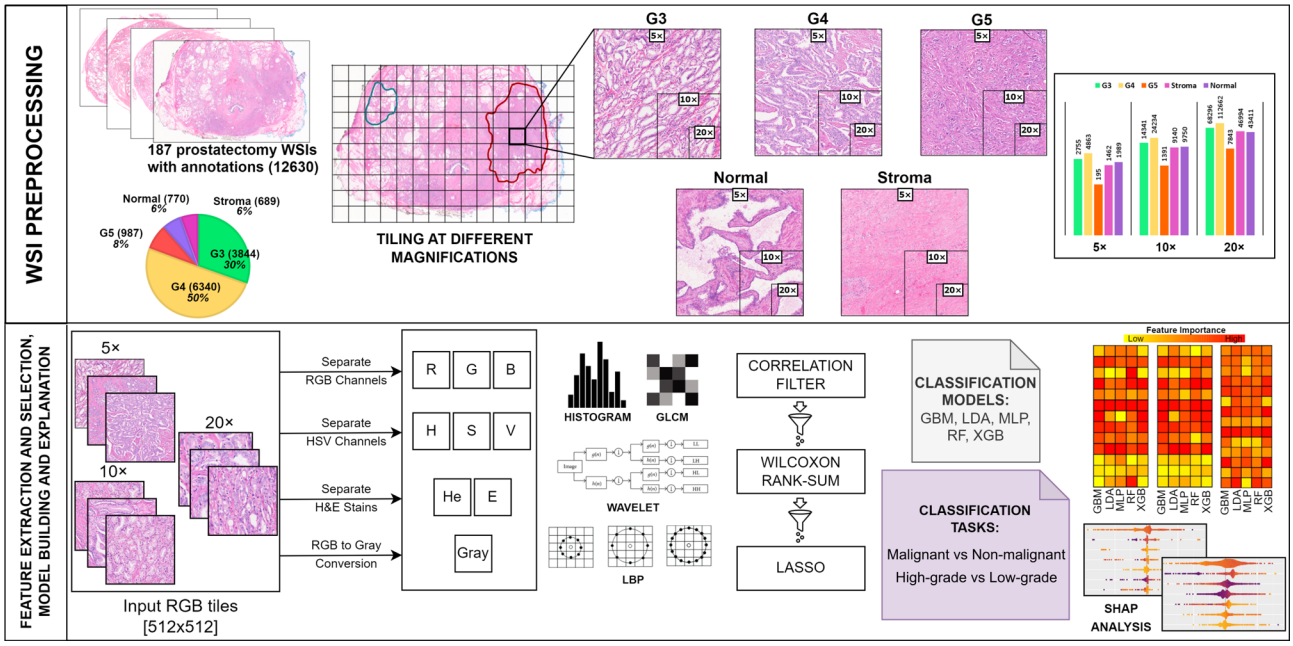

Fig. 2 Workflow of the pathomic approach implemented in the study. Starting from H&E prostatectomy WSIs, the approach starts with the WSI tiling at three different magnification scales (5×, 10× and 20×). Pathomic features were then extracted from each tile and selected to train classification models

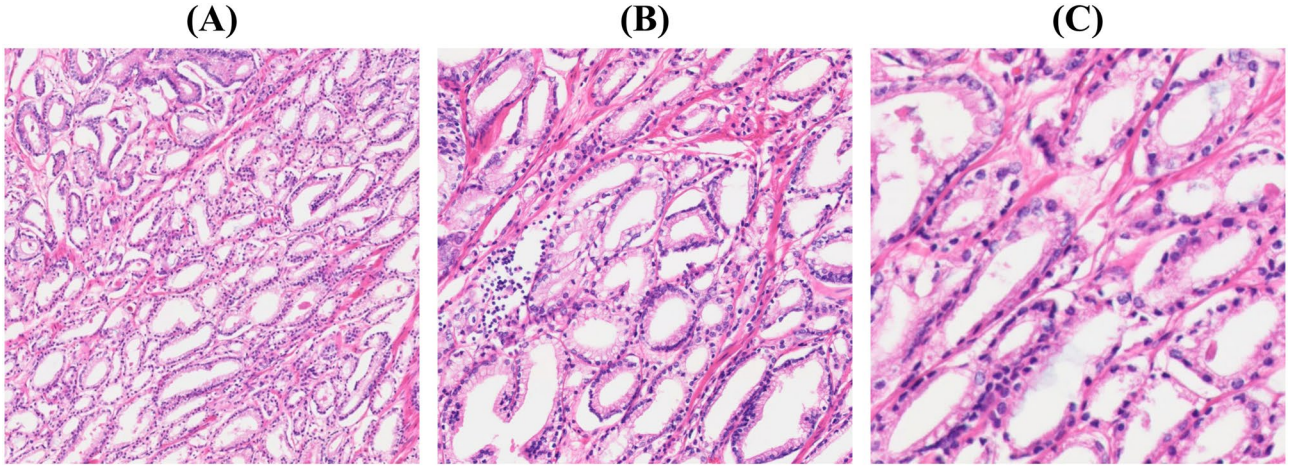

Fig. 3 Example of 512 × 512 tiles of a Gleason pattern 3. Gleason pattern 3 tile at 5× (**A**), 10× (**B**), and 20× (**C**) magnification

As suggested in Doyle et al. [21], histological features may be meaningful at different resolutions. Thus, image patches at different magnification scales were processed.

Three magnification levels, 5×, 10× and 20×, were investigated, and a patch size of 512 × 512 pixels for tiling without overlap (thus, the scales correspond to $1024 \times 1024 \mu m^2$, $512 \times 512 \mu m^2$ and $256 \times 256 \mu m^2$, respectively). The patches are cropped from the native 20× resolution of the core image, and lower scales (10× and 5×) are downsampled into patches of size 512 × 512 pixels (Fig. 3). The tiling algorithm that developed selects tiles from the images while obeying the following rules: at least 90% overlap with tissue annotations and tile overlap of 0%. ROIs that were too small to extract 5 tiles from were excluded.

Feature extraction

Four sets of descriptors were investigated: histogram, texture features computed in terms of gray level co-occurrence matrix (GLCM), wavelet and local binary pattern (LBP). All features were extracted from the original native RGB images converted to grayscale, as well as from the RGB and HSV color spaces. Additionally, features were derived from H and E stains separated through color deconvolution.

Tiles were processed using a custom, in-house MATLAB function to extract pathological features. The first

set was constructed using the gray level histogram of each image, yielding 18 histogram-based features.

The second set consisted of 14 Haralick features derived from the gray-level co-occurrence matrix (GLCM) as described by Haralick et al. [22] were calculated. The third set included 128 wavelet features obtained by applying a discrete wavelet transform with the Haar wavelet function to each pathology image, followed by the extraction of histogram and GLCM texture features from the wavelet-transformed images. Lastly, the fourth set comprised 59 uniform local binary pattern (LBP) features, extracted using MATLAB's extractLBPFeatures function [23, 24].

Overall, 219 features were extracted from 9 images, namely the RGB-to-grayscale conversion image, Red (R), Green (G), Blue (B), Hue (Hu), Saturation (S), Value (V), H stain (H), E stain (E), for a total of 1971 features for each tile, regardless of the associated magnification scale (5×, 10×, 20×). Refer to Supplementary Materials Sect. 2 for the complete feature glossary and information on feature calculations.

Feature selection

Features selection was conducted using a correlation filter that leverages the absolute values of pairwise Spearman's correlation (ρ) coefficient to minimize feature redundancy and prevent model overfitting. The threshold for ρ was set to 0.9. When two features had $\rho > 0.9$, the function evaluated the mean absolute correlation of each variable, removing the one with the largest mean absolute correlation. Further feature reduction was achieved through a univariate analysis using a nonparametric Wilcoxon rank-sum test to assess their statistical significance in relation to the binary outcome, depending on the classification task being investigated. Features with a p-value < 0.05 were retained. In order to account for multiple comparisons and control the false discovery rate (FDR), p-values were also adjusted using the Benjamini-Hochberg method [25]. The optimal feature set from these remaining features was then identified using the least absolute shrinkage and selection operator (LASSO) algorithm, which selects the most outcome-related features by minimizing the number of non-zero elements to ensure a unique solution. The shrinkage parameter lambda in the LASSO algorithm was chosen based on the smallest misclassification error in a 5-fold cross-validation. The selected features were then used to construct pathomic-based prediction models. These feature selection procedures were implemented using R studio software, version 4.0.2 (http://www.r-project.org/).

Data processing

Z normalization was performed on the training set prior model training and applied to the test set.

For binary classification, tissue tiles were grouped into distinct categories: G3, G4, and G5 tiles were classified as "Malignant," while normal and stroma tissue tiles were categorized as "Non-malignant." Additionally, G4 and G5 tissue tiles were combined into a "High-grade" group, with G3 designated as "Low-grade". The following classification tasks were investigated: Malignant vs. Non-malignant, and Low-grade vs. High-grade. These tasks were conducted at each magnification level, resulting in a total of six classification tasks.

Model building and performance evaluation

Before data processing, data were randomly split into training (80%) and testing sets (20%) but stratified to ensure that the ratio of positive samples to negative samples was the same between the training and test set.

This strategy maintained the proportional distribution of classes (positive or negative) in both the training and testing sets, ensuring an accurate representation of the various sample categories.

Additionally, to preserve the individuality of each patient in the training and testing data, an additional step was taken for balancing. While each patient may contribute multiple tiles to the overall dataset, it was ensured that patient IDs did not appear simultaneously in both the training and testing sets. This approach was implemented to prevent any potential loss of model generalization due to the simultaneous presence of the same patients in both sets, thus ensuring a fair and balanced representation of data diversity in the model training and evaluation process.

Five different models were used to evaluate the discrimination power of pathomic features: Gradient Boosting Machine (GBM), Linear Discriminant Analysis (LDA), Random Forest(RF), Multi-layer Perceptron (MLP), eXtreme Gradient Boosting (XGB). The Random forests were constructed using the ranger R package [26].

K-Fold patient-level cross-validation with $K = 5$ was applied to the training data set for model selection (features-selection, normalization and hyperparameter tuning). In this setup, folds were defined at the patient level to ensure that validation was always performed on patients not seen during training. To address the imbalance between positive and negative conditions, a downsampling strategy was employed, randomly reducing the majority class size to match or approximate the minority class size.

Receiver operating characteristic (ROC) curve analysis was applied to evaluate the predictive performance of the 6 classification models. The area under the receiver operating curve (AUC-ROC) was calculated and accuracy, sensitivity or recall, specificity, precision, and F1-score were employed to evaluate the diagnostic performance of resulting models in the hold-out test set. Confusion

matrices were computed on the test set to analyze model-specific misclassification patterns.

To further investigate the sources of prediction errors, a qualitative evaluation was performed by visually inspecting representative tiles from each confusion matrix category -True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). For each classification task, model, and magnification level, four mosaics (one per category) were generated by selecting samples from the test set, enabling an intuitive exploration of the morphological characteristics underlying correct and incorrect predictions. The complete set of mosaics is provided in the Supplementary Material (Figs. S3-S32).

To quantitatively support the visual inspection of model predictions, misclassified tiles were further analyzed by comparing image-derived features between selected confusion matrix categories – specifically FN vs. TP and FP vs. TN – across models, classification tasks, and magnification levels. These comparisons were chosen to focus on the distinction between correctly and incorrectly classified tiles within each class. Pairwise Wilcoxon tests with FDR correction were used to identify statistically significant differences potentially associated with misclassification.

Models building and evaluation were implemented using Caret package in R studio software, version 4.0.2, downloadable from http://www.R-project.org.

Model explanation and visualization

SHapley Additive exPlanations (SHAP) method was used to investigate models' interpretability and identify which were the key determinant features impacting on the models accounting for the classification tasks [27].

SHAP method is based on game-theoretically optimal Shapley values [27] and addresses the black-box nature of machine-learning models, enabling feature prioritization for classification.

SHAP explains predictions by computing each feature's contribution to the outcome. Using coalitional game theory, it treats features as players in a game, where the prediction represents the "victory." Shapley values fairly allocate contributions by evaluating all possible feature combinations and determining each feature's marginal impact [28]. SHAP has several advantages, including illustrating the importance of the features and their impact on the overall prediction model and understanding the importance of individual features to the model output in a model-independent way [29].

For each classification task, the feature importance bar plot and summary plot were used to understand the relationship between the feature and the impact on the prediction. Together, these visualizations enhance interpretability by illustrating both the relevance and directional impact of the top features for each model and magnification setting.

Moreover, given the multiple classification tasks performed at different magnification scales, as well as the five different models employed to investigate, a summarised representation of the results concerning the feature importance analysis was added.

Results

The total number of tiles according to tissue compartments and magnification scales are shown in Table 1.

For the diagnosis task, selection of pathomic features returned 53, 41, and 38 features, respectively for 5×, 10× and 20× magnification. For the grading task, selection of pathomic features returned 59, 53, and 51 pathomic features, respectively for 5×, 10× and 20× magnification.

Tables 2 and 3, and 4 reported the predictions of the classification models, respectively for 5×, 10× and 20× magnification. Figure 4 show the ROC curves representing performances of each classifier on the test set, at each magnification. Very high AUC values were reached for the Malignant vs. Non-malignant classification task ($0.97 \leq AUC \leq 0.99$). Lower AUC values for tasks involving the low-grade vs. high-grade classification, with maximum AUC values between 0.71 and 0.73 reached for MLP and LDA models, both for 5×, 10× and 20× magnification scales. Confusion matrices (Supplementary Figs. S1-S2) confirmed these trends, showing few misclassifications in the diagnosis task and a higher proportion of misclassifications in the grading task, particularly at lower magnifications.

Visual assessment of the mosaics of tile-level predictions (Supplementary Figs. S3–S32) offered further interpretability by highlighting recurring morphological trends across correctly and incorrectly classified tiles. For instance, some misclassified samples in the diagnostic task appeared to exhibit lower tissue density or

Table 1 Number of tiles according to tissue compartments and magnification scale

		Magnification		
		5×	10×	20×
Tissue Compartment	G3	2755	14,341	68,296
	G4	4863	24,234	112,662
	G5	195	1391	7843
	Normal	1462	9140	46,994
	Stroma	1989	9750	43,411

Table 2 Average prediction performance of different pathomic models for the two different classification tasks in the test set at 5× magnification level. Abbreviations: auc = area under the receiver operating characteristic curve; ci = confidence interval; sens = sensitivity; spec = specificity; prec = precision; ppv = positive predictive value; npv = negative predictive value; bacc = balanced accuracy; GSc = G-Score; gbm = gradient boosting machine; lda = linear discriminant analysis; MLP = Multi-layer perceptron; rf = random forest; xgb = extreme gradient boosting

	Model	AUCROC (CImin-CImax)	Accuracy (CImin-CImax)	Sens/Recall (CImin-CImax)	Spec (CImin-CImax)	Prec/PPV (CImin-CImax)	NPV (CImin-CImax)	F1	BAcc	G-score
Malignant vs. Non-Malignant	**GBM**	0.9674 (0.9612–0.9735)	0.9055 (0.8939–0.9162)	0.8687 (0.8438–0.8902)	0.9211 (0.9082–0.9324)	0.8242 (0.7973–0.8482)	0.9428 (0.9313–0.9524)	0.8459	0.8949	0.8945
	LDA	0.9695 (0.963–0.9761)	0.9198 (0.9089–0.9297)	0.9006 (0.8782–0.9193)	0.9279 (0.9154–0.9387)	0.8417 (0.8160–0.8645)	0.9564 (0.9461–0.9648)	0.8702	0.9143	0.9142
	MLP	0.9793 (0.9743–0.9844)	0.9388 (0.9291–0.9475)	0.9166 (0.8956–0.9337)	0.9483 (0.9374–0.9573)	0.8830 (0.8596–0.9029)	0.9639 (0.9545–0.9714)	0.8995	0.9324	0.9323
	RF	0.9654 (0.9592–0.9715)	0.8945 (0.8823–0.9057)	0.8748 (0.8504–0.8958)	0.9075 (0.8937–0.9197)	0.8011 (0.7736–0.8260)	0.9445 (0.9331–0.9541)	0.8306	0.8864	0.8861
	XGB	0.9717 (0.9656–0.9778)	0.9198 (0.9089–0.9297)	0.8883 (0.8649–0.9082)	0.9331 (0.9210–0.9435)	0.8498 (0.8242–0.8722)	0.9515 (0.9408–0.9603)	0.8686	0.9107	0.9105
High-grade vs. Low-grade	**GBM**	0.6461 (0.62-0.6723)	0.6253 (0.6033–0.647)	0.5500 (0.5100-0.5894)	0.6594 (0.6335–0.6844)	0.4220 (0.3878–0.4569)	0.7642 (0.7387–0.7879)	0.4776	0.6047	0.6022
	LDA	0.7321 (0.7078–0.7563)	0.6829 (0.6616–0.7037)	0.6383 (0.5991–0.6758)	0.7031 (0.6779–0.7271)	0.4929 (0.4579–0.5280)	0.8113 (0.7877–0.8329)	0.5563	0.6707	0.6699
	MLP	0.7287 (0.7033–0.7541)	0.6892 (0.668–0.7098)	0.6100 (0.5704–0.6482)	0.7249 (0.7003–0.7483)	0.5007 (0.4645–0.5368)	0.8043 (0.7809–0.8258)	0.5500	0.6675	0.6650
	RF	0.6255 (0.598–0.6531)	0.6196 (0.5975–0.6414)	0.4800 (0.4403-0.5200)	0.6586 (0.6327–0.6836)	0.3887 (0.3542–0.4242)	0.7369 (0.7111–0.7612)	0.4723	0.5996	0.5973
	XGB	0.6617 (0.6356–0.6878)	0.6388 (0.6169–0.6603)	0.5567 (0.5167–0.5959)	0.6760 (0.6503–0.7006)	0.4372 (0.4024–0.4726)	0.7713 (0.7463–0.7945)	0.4897	0.6163	0.6134

ambiguous glandular architecture, while in the grading task, increased morphological heterogeneity was more commonly observed in tiles associated with misclassified samples.

Pairwise comparisons of image-derived features between correctly and incorrectly classified tiles (TP vs. FN and TN vs. FP) revealed several statistically significant differences after FDR correction. These differences involved features related to texture, local intensity variation, and multi-scale structural information, including GLCM, LBP, and wavelet-based descriptors. A summary of significant features across comparisons is reported in Supplementary Sect. 5.

Concerning model explanation and visualisation using SHAP, the top 10 most important features responsible for each model, at each magnification scale, for each classification task are reported in Figs. 5 and 6 through side-by-side feature importance bar and summary plots. Moreover, a summarised representation of the most important features according to the SHAP results are reported in Figs. 7 and 8. Heatmaps representing feature importance were built for each model, at each magnification scale, for each of the four classification tasks.

For larity, and to improve the readability of the following text, feature names are abbreviated while retaining key information on the image channel, transformation type, and statistical metric. These abbreviations are used consistently throughout the rest of the Results and Discussion sections. A complete legend mapping each abbreviation to its full feature name is provided in Table 5.

Out of the most important features responsible for diagnosis task, R_or_fo_Median, R_wavLL_fo_SD, and S_or_fO_90P were found to be common in both 5×, 10× and 20× settings, regardless of the resampling technique. Feature H_wavHL_glcm_InfoCorrII was found to be among the most important features for both 10× and 20× classification tasks. Of note, E_wavLL_glcm_DV was among the most important features at 5× magnification, regardless of the resampling technique employed. The same happened for features H_wavHL_glcm_DV and H_wavLH_glcm _SV for the 10× classification task.

Out of the most important features responsible for grading task, S_wavLH_glcm _Var features were found to be the most common in settings 5× and 10×. Feature S_wavHL_glcm_ H_wavLL_fo_IQR were found to be the most common in 10× and 20× classification tasks. Of note, H_wavHL_glcm_SV and B_or_glcm_Var _glcm were among the most important features at 5× and at 10× magnification scales, respectively, regardless of the resampling technique. B_or_fo_90P, G_lbp_hf_Var7, and G_lbp_hf_Var9 were among the most important features at 20× magnification.

Discussion

Integrating AI into clinical practice is limited by challenges in interpreting features and their relevance to prediction models. This study aimed to identify key features for distinguishing malignant from non-malignant prostate tissue and classifying Gleason patterns using a multi-scale, multi-channel pathomic approach on H&E-stained images.

A comprehensive evaluation of the model's performance in these tasks was pursued as a secondary objective.

To address the critical issue of interpretability, SHAP was employed given its capability of illustrating the importance of the features and their impact on the overall prediction model and understanding the importance of individual features to the model output [27, 28]. By combining SHAP and pathomic, the aim of the study was to provide a transparent and clinician-friendly framework for PCa diagnosis and classification.

According to the obtained results, it is interesting to note that key features were found in common across different magnifications.

S_or_fo_90P and R_or_fo_Median were key determinant features impacting on models addressing for "Malignant vs Non-malignant" classification task, in both three magnification scales.

SHAP summary plots further reveal that H_wavHL_glcm_InfoCorrII has the greatest influence on distinguishing malignant from non-malignant patterns at 10× and 20× magnification.

Interestingly, features derived from wavelet transformations are consistently prominent in "High-grade vs Low-grade" classification task, confirming their significance in capturing texture and structural information from histopathological images and suggesting their relevance in characterizing tissue patterns associated with tumor grading [12, 30, 31]. LBP features also contributed to the classification, confirming their value in capturing tissue microstructure and texture [32, 33].

It is worth noting that also some histogram features from color channels of original images such as Ninetieth Percentile, Median, and Variance appeared as key features for diagnostic task. Colour histogram and wavelet features were also found to be relevant for PCa grading and diagnosis in study by Bhattacharjee et al. [30]. Moreover, Tabesh et al. [34], also found that color channel histograms were fairly effective in tumor/non-tumor classification. This is interesting given the simplicity of these features, which may reflect the overall intensity distribution within the tissue, possibly indicating differences in cellularity or tissue composition between malignant and non-malignant regions [35].

Common features across magnifications indicate their robustness as tissue structure descriptors, consistently

Table 3 Average prediction performance of different pathomic models for the two different classification tasks in the test set at 10× magnification level. Abbreviations: auc = area under the receiver operating characteristic curve; ci = confidence interval; sens = sensitivity; spec = specificity; prec = precision; ppv = positive predictive value; npv = negative predictive value; bacc = balanced accuracy; GSc = G-Score; gbm = gradient boosting machine; lda = linear discriminant analysis; MLP = Multi-layer perceptron; rf = random forest; xgb = extreme gradient boosting

	Model	AUCROC (CImin-CImax)	Accuracy (CImin-CImax)	Sens/Recall (CImin-CImax)	Spec (CImin-CImax)	Prec/PPV (CImin-CImax)	NPV (CImin-CImax)	F1	BAcc	G-score
Malignant vs. Non-Malignant	**GBM**	0.9987 (0.9984–0.9991)	0.9873 (0.9852–0.9891)	0.9829 (0.9789–0.9861)	0.9897 (0.9874–0.9916)	0.9819 (0.9779–0.9853)	0.9903 (0.9880–0.9921)	0.9824	0.9863	0.9863
	LDA	0.9919 (0.9901–0.9937)	0.9665 (0.9634–0.9694)	0.9270 (0.9195–0.9339)	0.9889 (0.9865–0.9909)	0.9794 (0.9750–0.9831)	0.9598 (0.9556–0.9636)	0.9525	0.9580	0.9575
	MLP	0.9985 (0.9978–0.9992)	0.9891 (0.9872–0.9907)	0.9843 (0.9805–0.9874)	0.9918 (0.9897–0.9934)	0.9855 (0.9818–0.9884)	0.9911 (0.9889–0.9928)	0.9849	0.9880	0.9880
	RF	0.9992 (0.9991–0.9994)	0.9876 (0.9856–0.9894)	0.9807 (0.9765–0.9842)	0.9861 (0.9835–0.9884)	0.9757 (0.9710–0.9796)	0.9890 (0.9866–0.9910)	0.9829	0.9869	0.9869
	XGB	0.9995 (0.9993–0.9997)	0.9901 (0.9883–0.9917)	0.9877 (0.9842–0.9904)	0.9915 (0.9894–0.9932)	0.9851 (0.9814–0.9881)	0.9930 (0.9910–0.9945)	0.9864	0.9896	0.9896
High-grade vs. Low-grade	**GBM**	0.696 (0.6844–0.7076)	0.6461 (0.6355–0.6566)	0.6813 (0.6657–0.6964)	0.6178 (0.6033–0.6320)	0.5894 (0.5743–0.6044)	0.7064 (0.6919–0.7206)	0.6320	0.6495	0.6487
	LDA	0.7229 (0.7116–0.7342)	0.6637 (0.6532–0.6741)	0.6959 (0.6806–0.7109)	0.6378 (0.6235–0.6519)	0.6074 (0.5923–0.6224)	0.7226 (0.7083–0.7364)	0.6487	0.6669	0.6662
	MLP	0.6979 (0.6862–0.7095)	0.6412 (0.6305–0.6517)	0.7239 (0.7089–0.7384)	0.5746 (0.5599–0.5891)	0.5781 (0.5635–0.5926)	0.7210 (0.7059–0.7356)	0.6428	0.6492	0.6449
	RF	0.6813 (0.6696–0.6931)	0.6411 (0.6304–0.6516)	0.6327 (0.6167–0.6484)	0.6457 (0.6315–0.6598)	0.5899 (0.5742–0.6054)	0.6858 (0.6715–0.6998)	0.6149	0.6412	0.6412
	XGB	0.7167 (0.7054–0.728)	0.6569 (0.6464–0.6674)	0.6827 (0.6671–0.6978)	0.6362 (0.6219–0.6503)	0.6018 (0.5866–0.6168)	0.7134 (0.6991–0.7273)	0.6397	0.6594	0.6590

capturing broad architectural patterns or statistical properties regardless of magnification [5, 36, 37].

However, the presence of unique features associated with each magnification, such as E_wavLL_glcm_DV for the Malignant vs. Non-Malignant task at 5× and feature H_wavLH_glcm_SV at 10×, underscores that each magnification level provides distinct information about tissue morphology and texture [5, 36, 37].

Moreover, the presence of features that are important for individual ML models could stem from the inherent differences in how each ML algorithm processes and weighs features. This emphasizes the importance of utilizing diverse models to capture a comprehensive understanding of the data, as different algorithms may highlight distinct aspects of the underlying patterns present in the histopathological images [38].

It was foreseeable that the key features would be different between the diagnostic and grading tasks. Features important for low-high grade classification (e.g., nuclear morphology, mitosis, proliferation, complex tumor structures) could differ from those simply distinguishing malignant from non-malignant tissue (e.g., cell density, nuclear structure, necrosis, tissue organization), as they focus on more subtle shades of tumor pathology characterizing the transition from one Gleason grade to another [5].

Concerning the classification performances, it was foreseeable that G3, G4, and G5 patterns could be easily distinguished from patterns characterizing benign tissues, as also confirmed by the consistently high AUC values obtained for the PCa diagnosis task. Lower AUC values for tasks involving the low-grade vs. high-grade classification, with maximum AUC values between 0.71 and 0.73 reached for MLP and LDA models, both for 5×, 10× and 20× magnification scales.

To complement standard performance metrics and provide greater interpretability, confusion matrices and representative visualizations of tile-level predictions across models and magnifications were included. These analyses were aimed at highlighting common trends in model predictions and exploring possible sources of misclassification. In the diagnosis task, confusion matrices showed a very low rate of false predictions across all models, in line with the very high AUC values (0.97–0.99). Visual assessment of misclassified tiles suggested that errors often occurred in regions with reduced tissue content or less clearly defined glandular structures, which may affect feature extraction and model confidence. For the grading task, where performance was comparatively lower, confusion matrices revealed a higher frequency of misclassified cases, particularly at lower magnifications. The visual inspection of tiles suggested that samples misclassified by the models tended to exhibit more heterogeneous morphology or less distinctive glandular patterns.

This qualitative interpretation was supported by a complementary quantitative analysis, which revealed significant differences in image-derived features between correctly and incorrectly classified tiles (TP vs. FN and TN vs. FP). In particular, features capturing texture complexity, local intensity variation (e.g., GLCM, LBP), and multi-scale structural patterns (e.g., wavelet descriptors) were frequently associated with classification outcomes. These results, despite the variability in selected features across models and tasks, reinforce the notion that certain image characteristics may contribute to misclassification, especially in more challenging grading scenarios.

The obtained classification results align with Bhattacharjee et al. [30] who examined wavelet-based and color features to train MLP model to perform different classification tasks (Benign vs. Malignant, Benign vs. G3, Benign vs. G4, Benign vs. G5, and G3 vs. G4 + G5). Their results also revealed lower classification performance for the "G3 vs. G4 + G5" classification task (85% accuracy) with respect to the "Benign vs Malignant" task (95% accuracy). Similar results were found by Alexandratou et al. [39], who achieved an accuracy of 97.9% for tumour-non-tumour and a lower accuracy of 80.8% for low-high grade discrimination. Kim et al. performed texture analysis using the GLCM method and carried out classification using SVM and KNN models for multiple tasks. The highest accuracy of 90% was achieved for benign vs. grade 4 and 5 [40]. Xu et al. also presented an automatic LBP-based approach for PCa grading [9]. They obtained an accuracy of 77% and an AUC of 0.93 for a three-class Gleason Score classification task (Gleason score 6 vs. Gleason score 7 vs. Gleason score ≥ 8).

While these studies consistently support the relevance of texture and color features in PCa diagnosis and grading, and their reported performances provide valuable reference points, direct comparison remains partially limited by the heterogeneity of experimental settings, including differences in WSI preprocessing, feature extraction pipelines, feature types, and classification frameworks. Nevertheless, these results serve as an important benchmark, helping to contextualize the performance of the present work within the broader literature. Crucially, what distinguishes the present study is not only the use of a publicly available dataset, which ensures transparency and reproducibility, but also the integration of multiscale and multi-channel handcrafted features with SHAP-based interpretability, offering a comprehensive and transparent framework for PCa classification. To our knowledge, this is the first work to apply a pathomic approach combined with SHAP for both diagnosis (malignant vs. non-malignant) and Gleason pattern grading tasks on this dataset.

As anticipated, most recent literature on automatic Gleason Grading is based on DL approaches [14–17, 41].

Table 4 Average prediction performance of different pathomic models for the two different classification tasks in the test set at 20× magnification level. Abbreviations: auc = area under the receiver operating characteristic curve; ci = confidence interval; sens = sensitivity; spec = specificity; prec = precision; ppv = positive predictive value; npv = negative predictive value; bacc = balanced accuracy; GSc = G-Score; gbm = gradient boosting machine; lda = linear discriminant analysis; MLP = Multi-layer perceptron; rf = random forest; xgb = extreme gradient boosting

	Model	AUCROC (CImin-CImax)	Accuracy (CImin-CImax)	Sens/Recall (CImin-CImax)	Spec (CImin-CImax)	Prec/PPV (CImin-CImax)	NPV (CImin-CImax)	F1	BAcc	G-score
Malignant vs. Non-Malignant	**GBM**	0.9952 (0.9947–0.9957)	0.973 (0.9716–0.9743)	0.9698 (0.9672–0.9722)	0.9744 (0.9728–0.9759)	0.9454 (0.9420–0.9486)	0.9860 (0.9848–0.9872)	0.9575	0.9721	0.9721
	LDA	0.986 (0.985–0.9869)	0.9509 (0.949–0.9526)	0.9201 (0.9160–0.9240)	0.9649 (0.9630–0.9667)	0.9230 (0.9190–0.9269)	0.9635 (0.9616–0.9653)	0.9216	0.9425	0.9422
	MLP	0.9965 (0.9962–0.9969)	0.9738 (0.9724–0.9751)	0.9743 (0.9718–0.9765)	0.9736 (0.9719–0.9752)	0.9440 (0.9406–0.9473)	0.9881 (0.9869–0.9891)	0.9589	0.9739	0.9739
	RF	0.9972 (0.997–0.9974)	0.9723 (0.9709–0.9736)	0.9691 (0.9664–0.9715)	0.9738 (0.9721–0.9753)	0.9441 (0.9406–0.9473)	0.9857 (0.9844–0.9868)	0.9564	0.9714	0.9714
	XGB	0.9981 (0.9979–0.9983)	0.982 (0.9809–0.9831)	0.9808 (0.9787–0.9827)	0.9826 (0.9812–0.9838)	0.9626 (0.9597–0.9653)	0.9911 (0.9901–0.9920)	0.9716	0.9817	0.9817
High-grade vs. Low-grade	**GBM**	0.6881 (0.6826–0.6936)	0.6114 (0.6064–0.6163)	0.7110 (0.7032–0.7187)	0.5568 (0.5506–0.5631)	0.4676 (0.4607–0.4745)	0.7788 (0.7725–0.7849)	0.5641	0.6339	0.6292
	LDA	0.7112 (0.7057–0.7166)	0.6207 (0.6157–0.6256)	0.7154 (0.7076–0.7230)	0.5688 (0.5626–0.5750)	0.4759 (0.4690–0.4829)	0.7850 (0.7789–0.7910)	0.5716	0.6421	0.6379
	MLP	0.7113 (0.7058–0.7168)	0.6206 (0.6156–0.6255)	0.7420 (0.7344–0.7494)	0.5541 (0.5478–0.5604)	0.4766 (0.4698–0.4835)	0.7969 (0.7907–0.8029)	0.5804	0.6480	0.6412
	RF	0.6716 (0.666–0.6772)	0.6144 (0.6095–0.6194)	0.6636 (0.6555–0.6716)	0.5875 (0.5813–0.5937)	0.4683 (0.4611–0.4754)	0.7614 (0.7552–0.7675)	0.5491	0.6256	0.6244
	XGB	0.6812 (0.6757–0.6866)	0.6175 (0.6126–0.6225)	0.6780 (0.6700-0.6859)	0.5844 (0.5782–0.5906)	0.4717 (0.4646–0.4788)	0.7683 (0.7621–0.7743)	0.5564	0.6312	0.6295

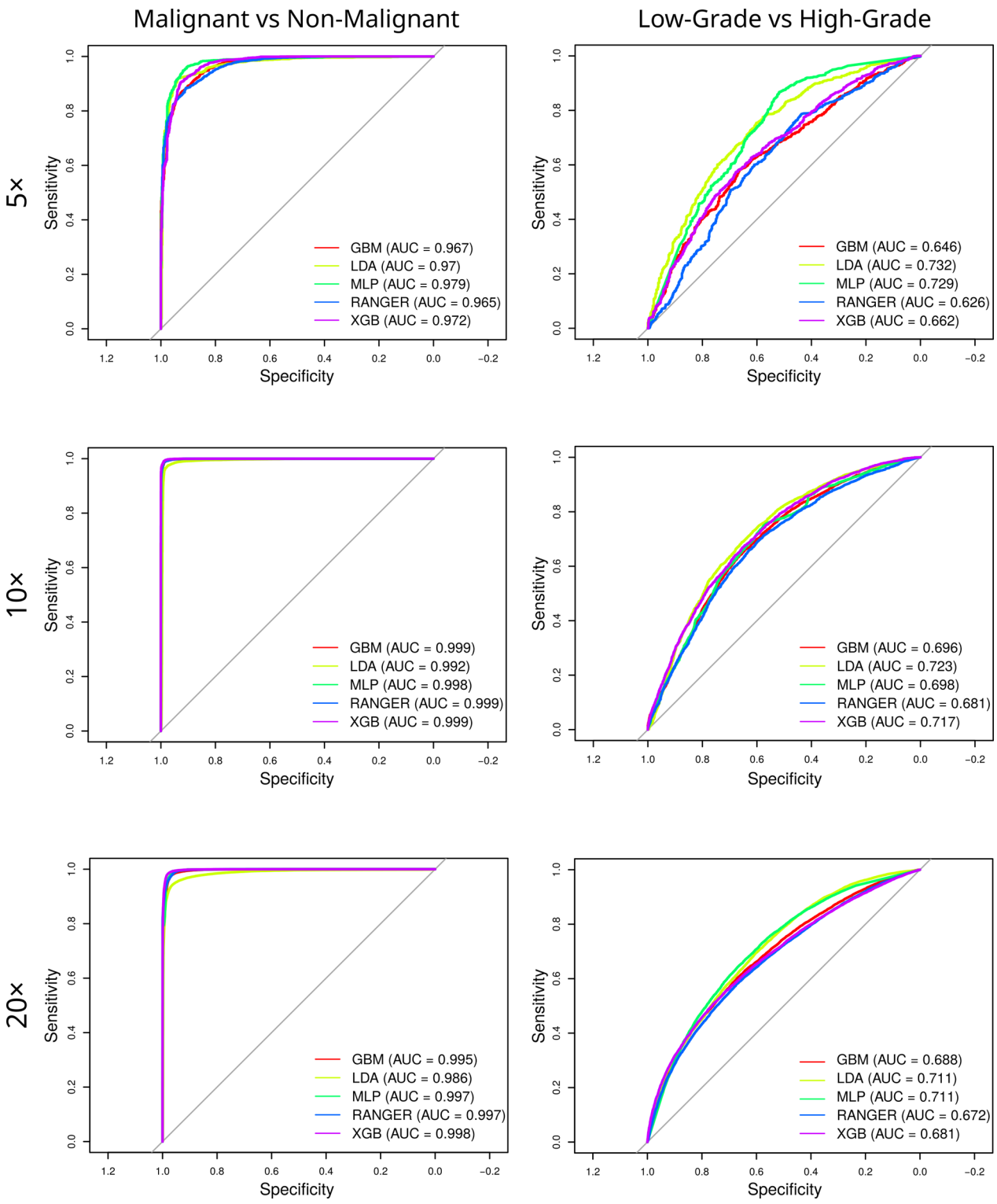

Fig. 4 Receiver Operating Characteristic (ROC) curves of prediction models. ROC curves are shown for each magnification scale (5×, 10×, and 20×, shown on the rows), and for each classification task ("Malignant vs. Non-Malignant" and "High-Grade vs. Low-Grade", shown on the columns). Each plot includes ROC curves for five classification models: Gradient Boosting Machine (GBM, red), Linear Discriminant Analysis (LDA, light green), Multilayer Perceptron (MLP, green), Random Forest (RANGER, blue), and Extreme Gradient Boosting (XGB, fuchsia)

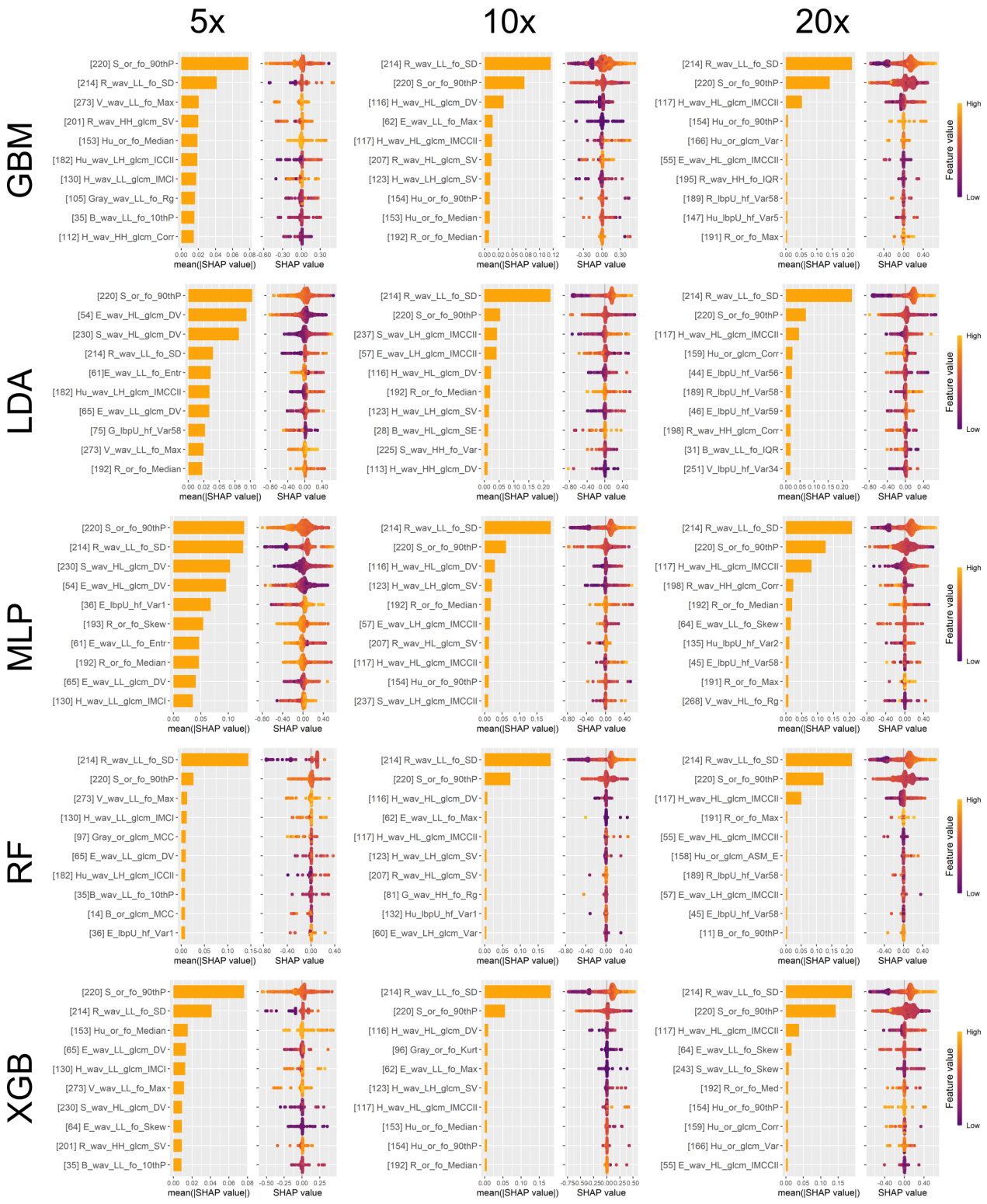

Fig. 5 SHAP feature importance and summary plots for the Malignant vs. Non-malignant classification task. Each subfigure displays the top ten most relevant features identified by SHAP values, shown side by side as a bar plot (left) and a summary plot (right). Rows correspond to classification models—Gradient Boosting Machine (GBM), Linear Discriminant Analysis (LDA), Multi-layer Perceptron (MLP), Random Forest (RF), and eXtreme Gradient Boosting (XGB)—while columns represent the three magnification levels (5×, 10×, 20×). Bar plots indicate the average absolute SHAP values of each feature, highlighting their overall importance in the classification process. Summary plots display the distribution of SHAP values for individual predictions, with each point representing a sample. Orange indicates high feature values, and purple indicates low values. A labeled colorbar or annotation indicates this color gradient in each plot

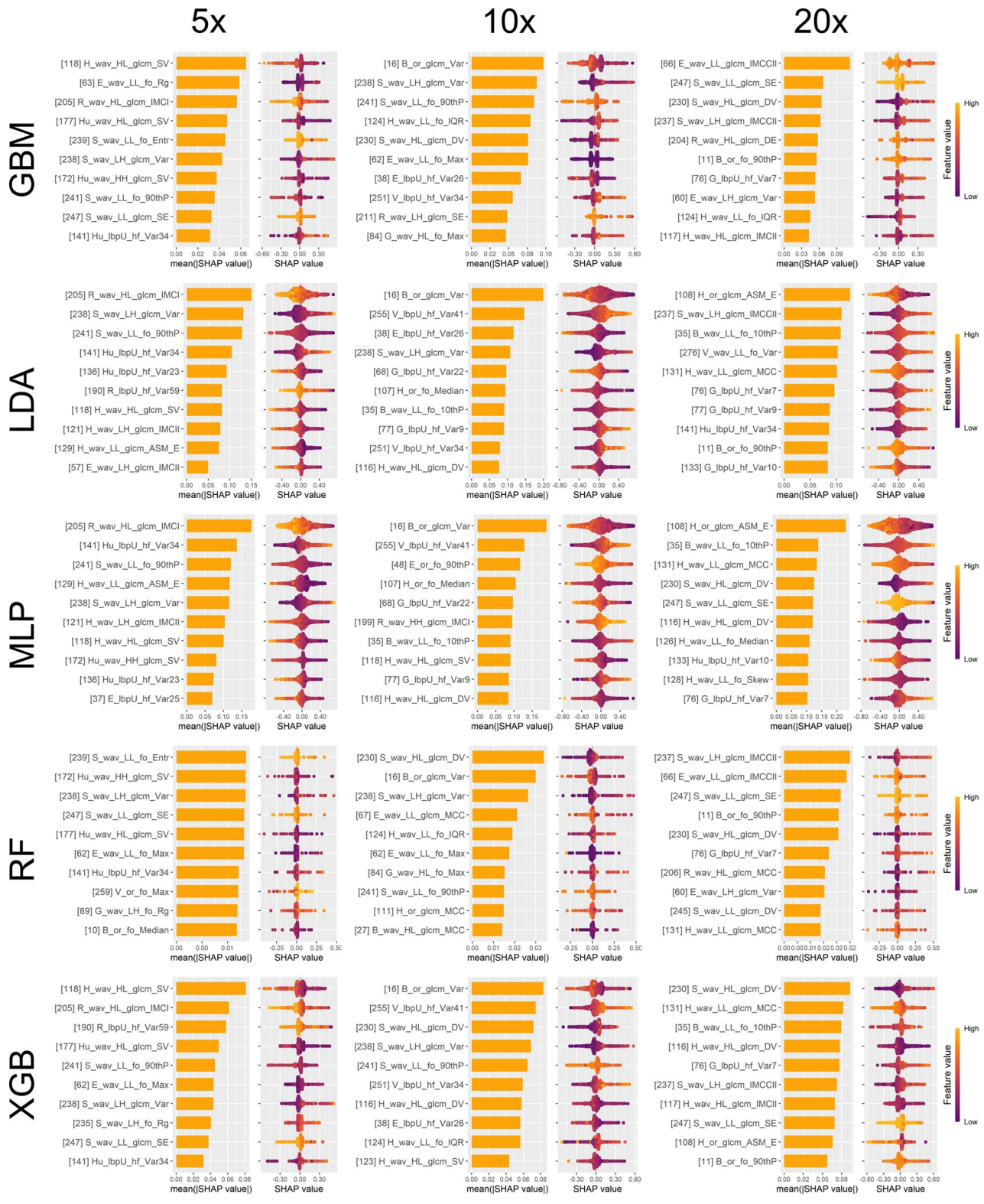

Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 SHAP feature importance and summary plots for the High-grade vs. Low-grade classification task. Each subfigure displays the top ten most relevant features identified by SHAP values, shown side by side as a bar plot (left) and a summary plot (right). Rows correspond to classification models—Gradient Boosting Machine (GBM), Linear Discriminant Analysis (LDA), Multi-layer Perceptron (MLP), Random Forest (RF), and eXtreme Gradient Boosting (XGB)—while columns represent the three magnification levels (5×, 10×, 20×). Bar plots indicate the average absolute SHAP values of each feature, highlighting their overall importance in the classification process. Summary plots display the distribution of SHAP values for individual predictions, with each point representing a sample. Orange indicates high feature values, and purple indicates low values. A labeled colorbar or annotation indicates this color gradient in each plot

Notably, Huo et al. released the image dataset employed in the present study and employed a multi-instance learning framework to analyse PCa WSIs, demonstrating the effectiveness of DL in capturing complex tissue patterns and achieving high accuracy in Gleason Grading cross-scanners. Their workflow integrates quality control (A! MagQC), cloud-based annotation (A! HistoClouds), and AI-assisted review (PAI), achieving high F1 scores in classifying Gleason patterns: 0.93 for G3, 0.84 for G4, 0.44 for G5, and 0.99 for non-malignant tissue. Notably, they also implemented image harmonization to reduce scanner variability, which improved the F1 score from 0.73 to 0.88 across external scanners [41].

Similarly, the PANDA Challenge [15] represents one of the largest and most comprehensive benchmarks for AI-based Gleason grading, involving over 10,000 biopsies across multiple international centers. The top-performing DL models achieved agreement levels comparable to expert pathologists (quadratically weighted $\kappa = 0.862–0.868$), confirming the scalability and robustness of DL approaches in multicenter settings.

While these studies highlight the strong performance of DL in prostate cancer grading, they typically rely on large datasets, high computational resources, and architectures that are often difficult to interpret. In contrast, the present study aims to provide a transparent and lightweight alternative, combining multiscale handcrafted features with SHAP-based interpretability.

Although it does not aim to compete directly with DL models, it proposes a viable and interpretable alternative suited for data-constrained or low-resource settings.

By adopting a pathomic framework based on handcrafted features and SHAP analysis, that emphasizes interpretability and applicability in data-constrained settings. By adopting a pathomic framework based on handcrafted features and SHAP analysis, the approach ensures transparency and offers direct associations between image features and tissue morphology—potentially providing novel insights for pathologists.

Texture-based features were chosen for their ability to represent patterns and structures in pathological images, capturing relevant information with less data than DL methods while maintaining accuracy [12]. This approach is suited for scenarios with limited datasets.

Unlike the majority of studies that are traditionally limited to specific feature subsets such as wavelet, color, or LBP [30, 42, 43], an expansive feature space was explored, integrating multiple channels beyond grayscale-converted RGB images, and incorporating color features to align with human visual interpretation in pathology.

The investigation of first-order, GLCM, LBP, and wavelet features stems from their extensive use in literature for a wide range of applications and their potential to capture crucial aspects of PCa pathology [21, 44].

Histogram-based features capture texture and outliers, while GLCM and LBP analyze pixel variability, offering insights into glandular structure and local texture. Wavelet features, which improve classification by capturing texture at multiple scales, provide a comprehensive understanding of PCa [30, 44, 45].

The decision to perform analyses at multiple magnification scales is based on a thoughtful consideration of what can be perceived at each level [37]. For example, at middle magnification levels (such as 5–10×) it is possible to distinguish between glands, while at the highest ones (such as 20–40×) it is possible to better resolve cells [46].

Of note, some few approaches to integrate multiple magnification levels have recently been explored [17, 37].

This departure from focusing on specific scales enriches the diagnostic landscape, providing enhanced granularity crucial for nuanced Gleason grading.

Despite the encouraging results, several limitations need to be addressed to ensure accurate interpretation and guide future research.

The relatively small dataset (187 prostatectomies) may impact model generalizability, and larger, more diverse datasets are needed for further validation. The small sample size is partly due to challenges in digitizing WSIs, as WSI technology is still seen as an augmentation rather than a replacement for traditional microscopy, limiting DL approaches that require large datasets [47].

Another limitation of this study is the absence of external validation, which is essential to fully assess model generalizability in real-world settings. However, to mitigate overfitting and assess robustness, both the final evaluation and internal model selection were conducted under strict patient-level partitioning. Specifically, a 5-fold patient-level cross-validation was applied during training, ensuring that validation was always performed on patients unseen by the model. Final performance metrics were computed on a separate, patient-exclusive hold-out test set. This design reduces the risk of overfitting and better simulates deployment on unseen clinical cases,

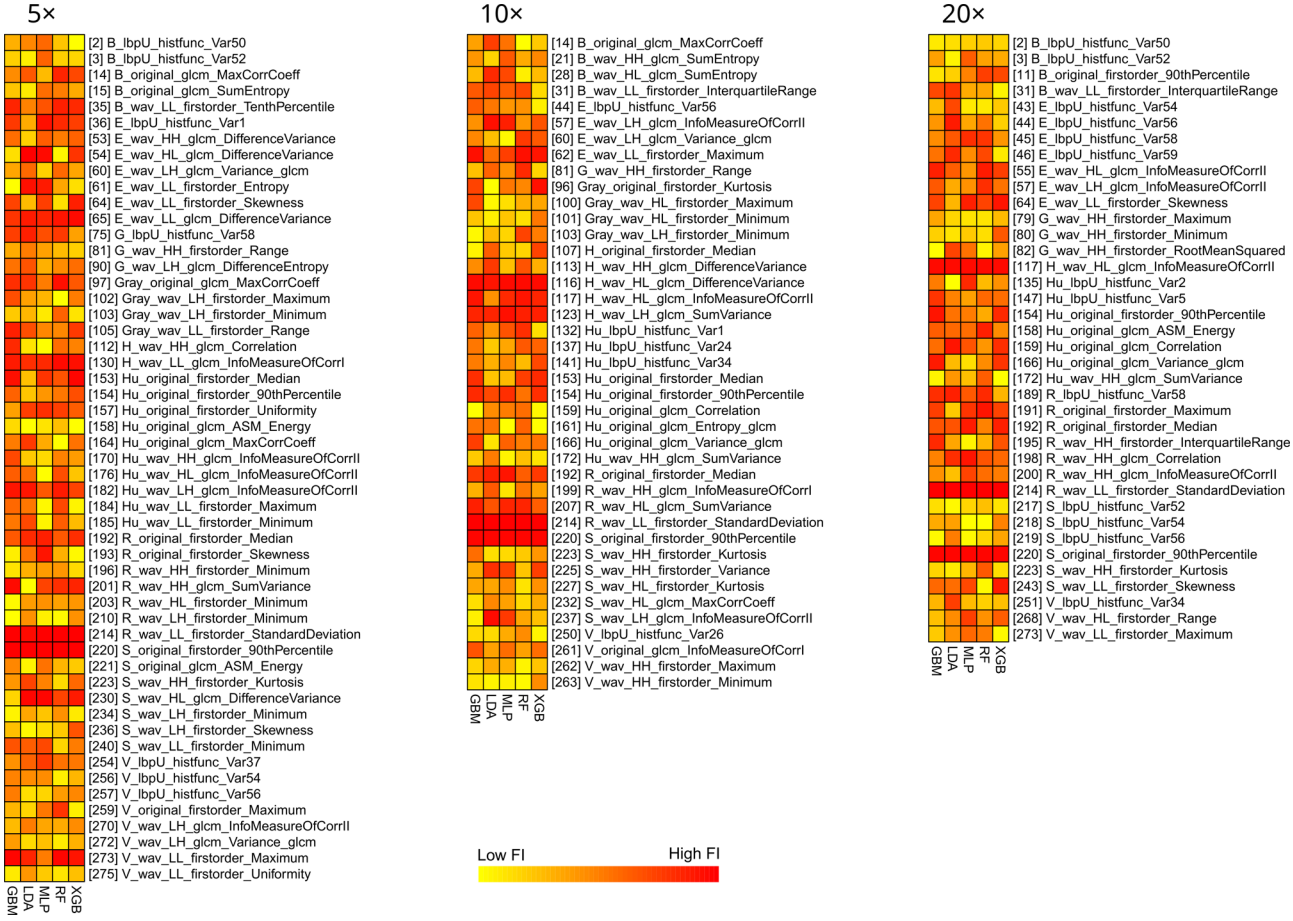

Fig. 7 Summarised representation of the most important features according to SHAP analysis for the "Malignant vs Non-Malignant" classification task. The summarised representation includes 5× (left), 10× (center), and 20× (right) magnification scale. Selected features contributing to the models were listed on the right. Each square of the heatmaps corresponds to a certain feature importance according to the colorbar. The colorbar is annotated to indicate that yellow = low feature importance, red = high importance

particularly important in tile-based pathology where intra-patient correlation can otherwise bias results.

Although the public dataset included duplicate acquisitions of the same prostatectomies on different scanners, these alternative scans could not be used for validation due to redundancy at the tissue level, which would artificially inflate performance.

Future work will focus on acquiring larger, multi-institutional datasets with independently digitized cases, to enable fully external validation and support clinical translation.

Another issue to be consider is the challenge of using color features due to high correlation between color channels [24]. To address this, a correlation filter was applied during feature selection to mitigate dependencies.

At last, the G5 pattern was not treated as a standalone class due to its rarity in the dataset and was combined with other groups (G4 for "High-Grade" and G3 + G4 for "Malignant"). This scarcity also posed a limitation in the study by Huo et al. [41], who emphasized the need for more G5 data for better evaluation. In the present study, G5 tiles originated from only 5 patients, making independent classification of this pattern unfeasible without incurring significant risks of class imbalance and overfitting. Although the number of tiles increases with magnification, this simply reflects finer sampling of the same few specimens rather than increased biological diversity.

Although data augmentation could in theory be used to boost the number of training samples [48], its utility in such low-patient contexts is questionable, as it may reinforce specimen-specific biases and further compromise model robustness.

Additionally, an intrinsic limitation of tile-based modeling in digital pathology must be acknowledged. Despite careful patient-level partitioning between training and testing sets, multiple tiles extracted from the same histological specimen remain inherently correlated. These intra-sample dependencies may introduce subtle forms of data leakage and compromise the reliability of the model,

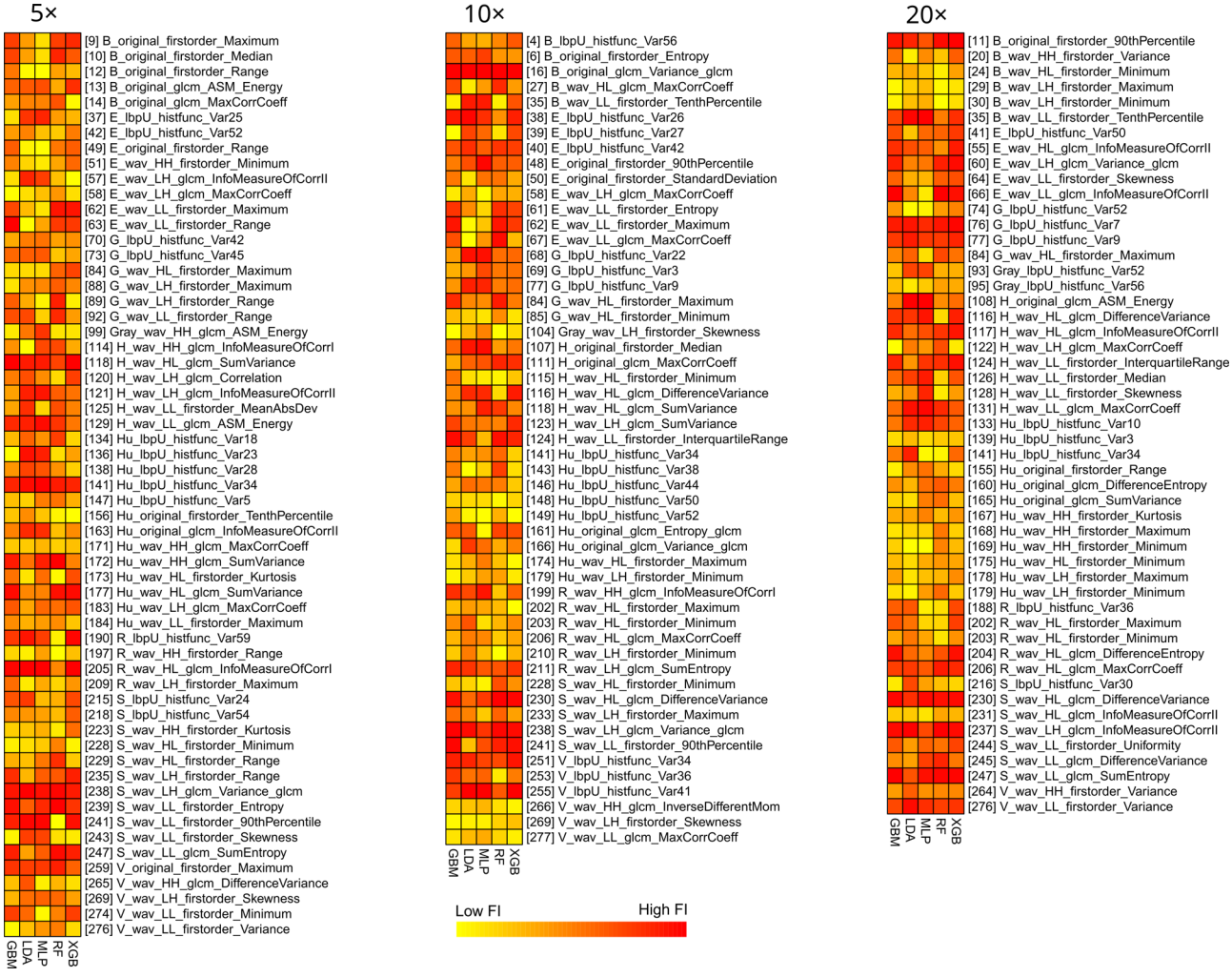

Fig. 8 Summarised representation of the most important features according to SHAP analysis for the "Gleason 4 + Gleason 5 vs Gleason 3" classification task. The summarised representation includes 5× (left), 10× (center), and 20× (right) magnification scale. Selected features contributing to the models were listed on the right. Each square of the heatmaps corresponds to a certain feature importance according to the colorbar. The colorbar is annotated to indicate that yellow = low feature importance, red = high importance

Table 5 Abbreviated feature names and their corresponding full descriptors

Feature name	Feature abbreviation
B_original_firstorder_NinetiethPercentile	B_or_fo_90P
B_original_glcm_Variance	B_or_glcm_Var
E_wav_LL_glcm_DifferenceVariance	E_wavLL_glcm_DV
G_lbpU_histfunc_Var7	G_lbp_hf_Var7
G_lbpU_histfunc_Var9	G_lbp_hf_Var7
H_wav_HL_glcm_DifferenceVariance	H_wavHL_glcm_DV
H_wav_HL_glcm_InformationMeasureOfCorrelationII	H_wavHL_glcm_InfoCorrII
H_wav_HL_glcm_SumVariance	H_wavHL_glcm_SV
H_wav_LH_glcm_SumVariance	H_wavLH_glcm _SV
H_wav_LL_firstorder_InterquartileRange	H_wavLL_fo_IQR
R_original_firstorder_Median	R_or_fo_Median
R_wav_LL_firstorder_StandardDeviation	R_wavLL_fo_SD
S_original_firstorder_NinetiethPercentile	S_or_fo_90P
S_wav_HL_glcm_DifferenceVariance	S_wavHL_glcm_DV
S_wav_LH_glcm_Variance	S_wavLH_glcm _Var

particularly for rare patterns like G5, where one or two patients may dominate the entire class distribution [49]. For these reasons, the G5 pattern was merged with other groups, prioritizing stability and interpretability of the classification task. Nonetheless, the biological and prognostic distinctiveness of G5 is fully recognized, and the development of future datasets with greater and more balanced G5 representation remains essential to enable dedicated modeling and refined risk stratification of high-grade prostate cancer.

To our knowledge, this study was the first to assess PCa diagnosis and grading by integrating pathomics with the SHAP method. Previous studies were performed on explainability of model based on radiomics, genomics and clinicopathologic features [50–52], as well as pathomic but applied to other cancer types [53, 54].

The strengths of the study lay also in its holistic and explicit approach, encompassing multiple magnification scales and channels for feature extraction. The transparent emphasis on interpretability addressed a critical concern in AI-based pathology, and the inclusion of an extensive feature space enhanced the versatility of the proposed methodology.

Future perspectives include exploring the fusion of information across magnification scales, mimicking pathologists' practices from low (e.g., 5×) to high (e.g., 20×) resolutions.

In this context, several existing articles explored the domain of multi-scale models employing DL methodologies. For example [55], introduces a pioneering cross-scale multi-instance learning algorithm that adeptly amalgamates multi-scale information and inter-scale relationships. Moreover, D'Amato et al. [37] also applied a multiple instance learning framework aiming at classifying histopathology images in both a single- and multi-scale setting, showing a consistent improvement in performance of the multi-scale models over single-scale ones.

Another avenue is comparing this approach with DL methods. While the present study focuses on hand-crafted features, future work could compare these with DL-based features for PCa Gleason Grade classification, potentially combining DL models with Deep SHAP for improved interpretability [27, 56, 57]. Moreover, future work could explore the application of emerging paradigms designed to address limited data availability and label noise, which are particularly relevant in histopathology. Semi-supervised learning approaches [58] could leverage unannotated regions of WSIs to improve generalization without requiring additional expert annotations. Few-shot learning [59] might support classification of rare patterns like Gleason 5, where training examples are extremely limited. Similarly, weakly supervised methods [60] could help exploit coarse or partial annotations—e.g., slide-level or region-level Gleason scores—thus reducing the annotation burden. Multimodal learning strategies [61] could enable the integration of histological features with clinical, genomic, or radiological data to improve risk stratification and disease characterization. Finally, the recent advances in large language and vision-language models (LLMs/VLMs) [62] may offer novel ways to incorporate prior medical knowledge and contextual reasoning into histopathological analysis.

Such directions are especially valuable for clinical translation, as they promise to increase robustness and reduce reliance on large, fully annotated datasets, which remain a bottleneck in computational pathology.

Conclusion

In conclusion, the study presented a pathomic-based approach to meet the challenge of the trade-off between performance and model interpretability for PCa diagnosis and grading. Key pathomic features were identified using various ML models with a wide range of features across magnification scales and color channels, showing promising performance. The approach offers clinician-friendly explanations and visualizations, advancing precision medicine for PCa and the real-world use of hand-crafted pathomic features.

The study also opened avenues for future research, emphasizing the need for ongoing refinement, validation on diverse datasets, and, by increasing the dimensionality of the dataset, the comparison with DL approaches.

Abbreviations

AI	Artificial Intelligence
WSI	Whole Slide Image
ML	Machine Learning
PCa	Prostate Cancer
MI-CLAIM	Minimum Information about Clinical AI Modeling
H&E	Hematoxylin and Eosin
GLCM	Gray Level Co-occurrence Matrix
LBP	Local Binary Pattern
RGB	Red Green Blue
HSV	Hue Saturation Value
ROC	Receiver Operating Characteristic
AUC-ROC	Area Under the Receiver Operating Curve
SHAP	SHapley Additive exPlanations
LDA	Linear Discriminant Analysis
MLP	Multi-layer Perceptron
RF	Random Forest
XGB	eXtreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01841-8.

Supplementary Material 1

Acknowledgements

The authors gratefully acknowledge the Computational and Quantitative Biology (CQB) Program of the University of Naples Federico II.

Author contributions
Conceptualization, V.B. and M.A.; Methodology, V.B. and M.V.; Data Curation: C.C., Visualization, F.I.; Writing-Original Draft Preparation, V.B. and M.V.; Writing-Review and Editing, C.C., F.I., M.A., and M.S.; Supervision, M.S. and M.A.

Funding
This work was partially funded by the Italian Ministry of Health ("Ricerca Corrente" project) and the European Union's Horizon Europe research and innovation program under grant agreement No. 101100633-EUCAIM.

Data availability
The original dataset analysed during this study are publicy available (https://aggc22.grand-challenge.org/AGGC22/). The processed datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 March 2025 / Accepted: 22 July 2025
Published online: 28 July 2025

References

1. Gupta R, Kurc T, Sharma A, Almeida JS, Saltz J. The emergence of pathomics. Curr Pathobiology Rep. 2019;7:73–84.
2. Azadi Moghadam P, Bashashati A, Goldenberg SL. Artificial intelligence and pathomics: prostate Cancer. Urol Clin North Am. 2024;51:15–26.
3. Bergengren O, Pekala KR, Matsoukas K, Fainberg J, Mungovan SF, Bratt O, et al. 2022 update on prostate Cancer epidemiology and risk Factors—A systematic review. Eur Urol. 2023;84:191–206.
4. Haffner MC, Zwart W, Roudier MP, True LD, Nelson WG, Epstein JI, et al. Genomic and phenotypic heterogeneity in prostate cancer. Nat Rev Urol. 2021;18:79–92.
5. Epstein JI. Prostate cancer grading: a decade after the 2005 modified system. Mod Pathol. 2018;31:47–63.
6. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A contemporary prostate Cancer grading system: A validated alternative to the Gleason score. Eur Urol. 2016;69:428–35.
7. Linkon AHMd, Labib MM, Hasan T, Hossain M, Jannat M-E-. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: an extensive study. Inf Med Unlocked. 2021;24:100582.
8. Short E, Warren AY, Varma M. Gleason grading of prostate cancer: a pragmatic approach. Diagn Histopathology. 2019;25:371–8.
9. Xu H, Park S, Hwang TH. Computerized classification of prostate Cancer Gleason scores from whole slide images. IEEE/ACM Trans Comput Biol Bioinf. 2020;17:1871–82.
10. Huo X, Ong K, Lau KW, Gole L, Young D, Tan CL et al. Comprehensive AI model development for gleason grading: from scanning, cloud-based annotation to pathologist-AI interaction [Internet]. In Review; 2022 Sep. Available from: https://www.researchsquare.com/article/rs-1917695/v1
11. Nguyen K, Sarkar A, Jain AK. Prostate cancer grading: use of graph cut and Spatial arrangement of nuclei. IEEE Trans Med Imaging. 2014;33:2254–70.
12. Mosquera-Lopez C, Agaian S, Velez-Hoyos A, Thompson I. Computer-Aided prostate Cancer diagnosis from digitized histopathology: A review on Texture-Based systems. IEEE Rev Biomed Eng. 2015;8:98–113.
13. Tolkach Y, Dohmgörgen T, Toma M, Kristiansen G. High-accuracy prostate cancer pathology using deep learning. Nat Mach Intell. 2020;2:411–8.
14. 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;13:2676.
15. Bulten W, Kartasalo K, Chen P-HC, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med. 2022;28:154–63.
16. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. Npj Digit Med. 2019;2:48.
17. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301–9.
18. Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: A powerful paradigm for scientific research. Innov. 2021;2:100179.
19. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4.
20. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X et al. A method for normalizing histology slides for quantitative analysis. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2009. pp. 1107–10.
21. Doyle S, Madabhushi A, Feldman M, Tomaszeweski J. A Boosting Cascade for Automated Detection of Prostate Cancer from Digitized Histology. In: Larsen R, Nielsen M, Sporring J, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006 [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006 [cited 2023 Mar 10]. pp. 504–11. Available from: http://link.springer.com/https://doi.org/10.1007/11866763_62
22. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;SMC–3:610–21.
23. Pietikäinen M, Hadid A, Zhao G, Ahonen T. Computer Vision Using Local Binary Patterns [Internet]. London: Springer London. 2011 [cited 2023 Dec 10]. Available from: http://link.springer.com/https://doi.org/10.1007/978-0-85729-748-8
24. Shu X, Song Z, Shi J, Huang S, Wu X-J. Multiple channels local binary pattern for color texture representation and classification. Sig Process Image Commun. 2021;98:116392.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc Ser B: Stat Methodol. 1995;57:289–300.
26. Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++and R. J Stat Softw. 2017;77:1–17.
27. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc. 2017 [cited 2024 Mar 14]. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
28. Molnar C. 9.6 SHAP (SHapley Additive exPlanations)| Interpretable Machine Learning [Internet]. [cited 2024 Apr 3]. Available from: https://christophm.github.io/interpretable-ml-book/shap.html
29. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des. 2020;34:1013–26.
30. Bhattacharjee S, Kim C-H, Park H-G, Prakash D, Madusanka N, Cho N-H, et al. Multi-Features classification of prostate carcinoma observed in histological sections: analysis of Wavelet-Based texture and colour features. Cancers. 2019;11:1937.
31. Lopez CM, Agaian S. A new set of wavelet- and fractals-based features for Gleason grading of prostate cancer histopathology images. Image Processing: Algorithms and Systems XI [Internet]. SPIE 2013 [cited 2024 Apr 3]. pp. 423–34. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8655/865516/A-new-set-of-wavelet--and-fractals-based-features/https://doi.org/10.1117/12.1000193.full
32. Farooq MT, Shaukat A, Akram U, Waqas O, Ahmad M. Automatic gleason grading of prostate cancer using Gabor filter and local binary patterns. 2017 40th International Conference on Telecommunications and Signal Processing (TSP) [Internet]. 2017 [cited 2024 Apr 3]. pp. 642–5. Available from: https://ieeexplore.ieee.org/document/8076065
33. Xu H, Park S, Hwang TH. Automatic Classification of Prostate Cancer Gleason Scores from Digitized Whole Slide Tissue Biopsies [Internet]. bioRxiv; 2018 [cited 2024 Apr 3]. p. 315648. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/315648v1
34. Tabesh A, Teverovskiy M, Pang H-Y, Kumar VP, Verbel D, Kotsianti A, et al. Multifeature prostate Cancer diagnosis and Gleason grading of histological images. IEEE Trans Med Imaging. 2007;26:1366–78.

35. Magi-Galluzzi C. Prostate cancer: diagnostic criteria and role of immunohistochemistry. Mod Pathol. 2018;31:12–21.
36. Humphrey PA. Histopathology of prostate Cancer. Cold Spring Harb Perspect Med. 2017;7:a030411.
37. D'Amato M, Szostak P, Torben-Nielsen B. A comparison between Single- and Multi-Scale approaches for classification of histopathology images. Front Public Health. 2022;10:892658.
38. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inf Decis Mak. 2019;19:281.
39. Alexandratou E, Atlamazoglou V, Thireou T, Agrogiannis G, Togas D, Kavantzas N, et al. Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading. Int J Comput Intell Bioinf Syst Biology. 2010;1:297–315.
40. Kim C-H, So J-H, Park H-G, Madusanka N, Deekshitha P, Bhattacharjee S, et al. Analysis of texture features and classifications for the accurate diagnosis of prostate Cancer. J Korea Multimedia Soc. 2019;22:832–43.
41. Huo X, Ong KH, Lau KW, Gole L, Young DM, Tan CL, et al. A comprehensive AI model development framework for consistent Gleason grading. Commun Med. 2024;4:84.
42. Tabesh A, Teverovskiy M, Tumor, Classification in Histological Images of Prostate Using Color Texture. 2006 Fortieth Asilomar Conference on Signals, Systems and Computers [Internet]. Pacific Grove, CA, USA: IEEE; 2006 [cited 2023 Mar 13]. pp. 841–5. Available from: http://ieeexplore.ieee.org/document/4176678/
43. Peyret R, Bouridane A, Khelifi F, Tahir MA, Al-Maadeed S. Automatic classification of colorectal and prostatic histologic tumor images using multiscale multispectral local binary pattern texture features and stacked generalization. Neurocomputing. 2018;275:83–93.
44. Kather JN, Weis C-A, Bianconi F, Melchers SM, Schad LR, Gaiser T, et al. Multi-class texture analysis in colorectal cancer histology. Sci Rep. 2016;6:27988.
45. Roberti de Siqueira F, Robson Schwartz W, Pedrini H. Multi-scale Gray level co-occurrence matrices for texture description. Neurocomputing. 2013;120:336–45.
46. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. Histopathology. 2012;61:1–9.
47. Atallah NM, Toss MS, Verrill C, Salto-Tellez M, Snead D, Rakha EA. Potential quality pitfalls of digitalized whole slide image of breast pathology in routine practice. Mod Pathol. 2022;35:903–10.
48. Hassell LA, Forsythe ML, Bhalodia A, Lan T, Rashid T, Powers A et al. Toward optimizing the impact of digital pathology and augmented intelligence on issues of diagnosis, grading, staging and classification. Mod Pathol. 2025;100765.
49. Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C et al. AI Slipping on Tiles: Data Leakage in Digital Pathology. In: Del Bimbo A, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, editors. Pattern Recognition ICPR International Workshops and Challenges. Cham: Springer International Publishing; 2021. pp. 167–82.
50. Suh J, Yoo S, Park J, Cho SY, Cho MC, Son H, et al. Development and validation of an explainable artificial intelligence-based decision-supporting tool for prostate biopsy. BJU Int. 2020;126:694–703.
51. Rodrigues A, Rodrigues N, Santinha J, Lisitskaya MV, Uysal A, Matos C, et al. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. Sci Rep. 2023;13:6206.
52. Ramírez-Mena A, Andrés-León E, Alvarez-Cubero MJ, Anguita-Ruiz A, Martinez-Gonzalez LJ, Alcala-Fdez J. Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. Comput Methods Programs Biomed. 2023;240:107719.
53. Altini N, Puro E, Taccogna MG, Marino F, De Summa S, Saponaro C, et al. Tumor cellularity assessment of breast histopathological slides via instance segmentation and pathomic features explainability. Bioengineering. 2023;10:396.
54. Zhan F, Guo Y, He L. NETosis Genes and Pathomic Signature: A Novel Prognostic Marker for Ovarian Serous Cystadenocarcinoma. J Digit Imaging Inform med [Internet]. 2024 [cited 2025 Jan 21]. Available from: https://link.springer.com/https://doi.org/10.1007/s10278-024-01366-6
55. Deng R, Cui C, Remedios LW, Bao S, Womick RM, Chiron S et al. Cross-scale Multi-instance Learning for Pathological Image Diagnosis [Internet]. arXiv. 2024 [cited 2024 Apr 3]. Available from: http://arxiv.org/abs/2304.00216
56. Tortora M, Cordelli E, Sicilia R, Nibid L, Ippolito E, Perrone G, et al. Radio-Pathomics: multimodal learning in Non-Small cell lung Cancer for adaptive radiotherapy. IEEE Access. 2023;11:47563–78.
57. Huang X, Li Z, Zhang M, Gao S. Fusing hand-crafted and deep-learning features in a convolutional neural network model to identify prostate cancer in pathology images. Front Oncol. 2022;12:994950.
58. Chen J, Zhang J, Debattista K, Han J. Semi-Supervised unpaired medical image segmentation through Task-Affinity consistency. IEEE Trans Med Imaging. 2023;42:594–605.
59. Chen J, Chen C, Huang W, Zhang J, Debattista K, Han J. Dynamic contrastive learning guided by class confidence and confusion degree for medical image segmentation. Pattern Recogn. 2024;145:109881.
60. Chen J, Huang W, Zhang J, Debattista K, Han J. Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. IEEE Trans Image Process. 2025. PP.
61. Chen J, Li W, Li H, Zhang J. Deep Class-Specific Affinity-Guided Convolutional Network for Multimodal Unpaired Image Segmentation [Internet]. arXiv. 2021 [cited 2025 Jun 12]. Available from: https://arxiv.org/abs/2101.01513
62. Chen J, Li H, Zhang J, Menze B. Adversarial Convolutional Networks with Weak Domain-Transfer for Multi-Sequence Cardiac MR Images Segmentation [Internet]. arXiv. 2019 [cited 2025 Jun 12]. Available from: https://arxiv.org/abs/1908.09298

Publisher's note