RESEARCH

Evaluation of automatic discrimination between benign and malignant prostate tissue in the era of high precision digital pathology

Yauheniya Zhdanovich[1]
, Jörg Ackermann[2]
, Peter J. Wild[4]
, Jens Köllermann[4]
, Katrin Bankov[4]
, Claudia Döring[4]
, Mike Wenzel[6]
, Benedikt Höh[6]
, Philipp Mandel[6]
, Thomas J. Vogl[3]
, Patrick Harter[5]
, Katharina Filipski[5]
, Ina Koch[2]
and Simon Bernatz[3,4,7*]

Abstract

Background: Prostate cancer is a major health concern in aging men. Paralleling an aging society prostate cancer prevalence increases emphasizing the need for efficient diagnostic algorithms.

Methods: Retrospectively, 106 prostate tissue samples from 48 patients (mean age, 66 +/- 6.6 years) were included in the study. Patients suffered from prostate cancer (n=38) or benign prostatic hyperplasia (n=10) and were treated with radical prostatectomy (RP) or Holmium laser enucleation of the prostate (HoLEP), respectively. We constructed tissue microarrays (TMAs) comprising representative malignant (n=38) and benign (n=68) tissue cores. TMAs were processed to histological slides, stained, digitized and assessed for the applicability of machine learning strategies and open source tools in diagnosis of prostate cancer. We applied the software QuPath to extract features for shape, stain intensity and texture of TMA cores for three stainings, H&E, ERG, and PIN-4. Three machine learning algorithms, neural network (NN), support vector machines (SVM), and random forest (RF), were trained and cross-validated with 100 Monte Carlo random splits into 70% training set and 30% test set. We determined AUC values for single color channels, with and without optimization of hyper-parameters by exhaustive grid search. We applied recursive feature elimination (RFE) to feature sets of multiple color transforms.

Results: Mean AUC was above 0.80. PIN-4 stainings yielded higher AUC than H&E and ERG. For PIN-4 (color transform saturation), NN, RF, and SVM revealed AUC of 0.92 ± 0.06, 0.91 ± 0.05, and 0.93 ± 0.04, respectively. Optimization of hyper-parameters improved the AUC only slightly by $0.01 - 0.02$. For H&E, feature selection resulted in no increase of AUC but to an increase of $0.02 - 0.05$ for ERG and PIN-4.

Conclusions: Automated pipelines may be able to discriminate with high accuracy between malignant and benign tissue. We found PIN-4 staining best suited for classification. Further bioinformatic analysis of larger data sets would be crucial to evaluate the reliability of automated classification methods for clinical practice and to evaluate potential discrimination of aggressiveness of cancer to path the way to automatic precision medicine.

Keywords: prostate cancer; prediction; quantitative features; statistical analysis; machine learning; neural networks; automated pipelines

Introduction

Prostate cancer (PCa) is the second most common cancer and the fifth leading cause of cancer death in men [1]. Incidence rates vary across regions and PCa is the most frequently diagnosed cancer in men in 112 of 185 countries of the world [1]. One established risk factor is an advancing age [1]. Due to the demographic development of an aging society we may expect an increasing PCa burden in the future [1]. Diagnosis of clinically significant prostate cancer is a challenging process. Most prostate cancers are slow-growing, a subset of prostate cancers has an aggressive clinical course and leads to death. Prostate cancer is usually suspected on the basis of screening procedures: digital rectal examination and/or prostate-specific antigen levels [2]. For definitive diagnosis, histopathological verification of PCa in prostate biopsy cores is required. Grading of PCa with the Gleason system is the strongest prognostic factor for clinical behaviour and treatment response [2, 3]. Computerization and the efficient addressing of crucial cancer care touchpoints along the patient clinical pathway are major goals of current studies in the field of artificial intelligence (AI) in medicine [4]. Clinical decision support systems aim to assist physicians and other specialists in the analysis of patient's data and diagnosis of diseases [5]. Quantitative imaging, machine learning (ML) algorithms, and AI have been proposed as potential solutions for assisting clinicians [6]. ML and AI have the potential to improve the accuracy and robustness of the diagnosis of PCa [7, 8]. Recent studies on PCa addressed the prediction of Gleason grade scores [9], the detection of PCa in biopsy specimen [10], the extraction of cancer stage from written reports in structured form [11], and the prediction of risk of PCa based on demographic characteristics [12]. Features extracted from digital images in pathology may have the potential to predict recurrence in PCa patients after surgery [13]. ML and AI have been used for cancer detection and grading based on whole image analysis in prostate biopsies [14, 15, 10, 16, 17, 18]. For the classification of benign and malignant tissues, multi-view boosting methods have been proposed [19]. The results have been compared to single-view classification and have reached a high area under the curve (AUC) score of 0.98 [19]. For a review of applications of deep learning to cancer detection, we refer to Pantanowitz *et al.* [20] and literature cited therein. Application of ML approaches may help in assisting physicians in the examination and prioritization of patient's data, for a discussion, we refer to Bulten *et al.* [21].

The ground truth is usually based on visual inspection, evaluation and classification by expert pathologists. Besides H&E images additional immunohistochemical workup can aid the urologic subspecialist in identifying and classifying cancer, e.g., see [8, 22]. Researchers in the field of digital pathology have put a considerable effort in the development and evaluation of methods especially for H&E images. Despite the well known advantages of immunohistochemical stainings in daily standard of care clinical pathology, its possible additive prediction power is only scarcely studied and evaluated. Exploring the suitability of immunohistochemical stainings for automated classification tasks and machine learning models may provide an enhanced possibility for high precision digital pathology in the automatic classification of prostate cancer.

In this work, we considered the staining methods, H&E, ERG, and PIN-4. H&E is a standard staining among others for cancer diagnosis [23]. ERG expression is a potential biomarker to predict the aggressiveness of prostate carcinoma with potential prognostic impact [24, 25]. PIN-4 is a cocktail of multiple markers and may help to distinguish between high-grade prostatic intraepithelial neoplasia and adenocarcinoma [26, 27]. Sabata *et al.* have investigated the identification and classification of glands in a whole slide image of PIN-4 stained prostate needle biopsy [27]. To our knowledge, no study has considered the AI based prediction capability of PCa comparing the three commonly performed stainings: H&E, ERG, and PIN-4.

We retrospectively studied 106 tissue cores (malignant, n=38; benign, n=68) that we stained with three different methods, i.e., H&E, ERG, and PIN-4. Our purpose was to evaluate the suitability of basic image features based on the intensity distribution and the texture of the stained tissue cores to automatically differentiate between PCa tissue and benign tissue. We applied the open source software QuPath (version 0.2.0) [28] for segmentation and feature extraction. We evaluated the prediction power of the features with three standard ML methods, i.e. neural network (NN), support vector machines (SVM) and random forest (RF) with and without optimization of hyperparameters and strategies for feature selection, e.g. recursive feature elimination (RFE) [29]. We proposed a simple automated approach that might be feasible in clinical routine.

Results

We compared a group of malignant cores with a group of benign cores. If not indicated otherwise, p-values in

[*] Correspondence: Simon.Bernatz@kgu.de
[3] Department of Diagnostic and Interventional Radiology, Goethe University Frankfurt am Main, University Hospital Frankfurt, 60590 Frankfurt am Main, Germany
Full list of author information is available at the end of the article

the following were computed with the Wilcoxon-Mann-Whitney-U test and corrected for multiple testing by the Benjamini–Hochberg procedure. Table 1 shows the number of features with false discovery rates (FDR): $p \leq 0.05, p \leq 0.01$, and $p \leq 0.001$. Not surprisingly, none of shape features had FDR below 5%.

Table 1: Number of features for the H&E, ERG, and PIN-4 staining. We applied the Wilcoxon-Mann-Whitney-U test with Benjamini–Hochberg correction for multiple testing and selected features with low false discovery rate, $p \leq 0.05, p \leq 0.01$, and $p \leq 0.001$.

significance	H&E	ERG	PIN-4
all	166	165	117
$p \leq 0.05$	114	128	76
$p \leq 0.01$	105	111	67
$p \leq 0.001$	92	93	43

We computed pairwise Pearson correlation coefficients of features and found features from different color transforms highly correlated. Due to highly correlated features from different color transforms, the selection of features with highest Gini score of the Wilcoxon-Mann-Whitney-U test was not a valuable strategy. To reduce the redundancy of features, we decided to consider, in the first step, only features of a single color transform for each stain. We compared the prediction power of features in the color transforms to select a representative color transform for each stain.

Exemplary, Figure 1 shows the boxplots of Gini coefficients of sets of features with $p \leq 0.05$ specific for individual color transforms and stainings H&E (top), ERG (middle), and PIN-4 (bottom). A Gini coefficient of one corresponds to perfect prediction power whereas a Gini coefficient of zero corresponds to no prediction power, i.e., random choices. The prediction power of features varies for the stains. For H&E, the prediction powers of standard color transforms, Red, Green, Blue, Saturation, Brightness, and OD Sum, were preferable high with median Gini coefficients of in round numbers 0.6. For stain specific colors Hematoxylin and Eosin, the median Gini coefficient drops to in round numbers 0.4. Residual has the lowest median Gini coefficient of in round numbers 0.3. For ERG, the median Gini coefficients are also preferable high. Differences in the color transforms are more pronounced than in H&E. Blue has a higher median Gini coefficient than Red and Green. Note that, counterstaining with hematoxylin produces blue-purple signal for cell nuclei. High concentration of the protein ERG would manifest in a brownish nuclear signal, i.e., a high value of Red and Green. Surprisingly, the signal of counter-staining with hematoxylin has a higher median prediction power than the signal of ERG itself. Brightness has a higher prediction power than Saturation. Compared to ERG and H&E, the features of PIN-4 have rather low medians of Gini coefficient. An exceptional high Gini score of 0.865 gives, however, Maximum Saturation.

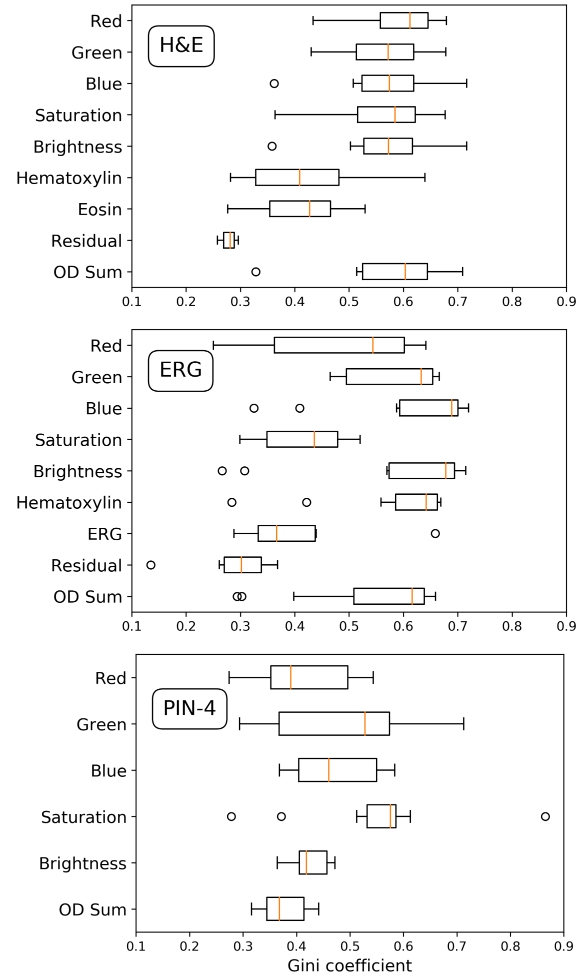

Figure 1: Prediction power of sets of features with $p \leq 0.05$ for H&E staining (top), ERG staining (middle), and PIN-4 staining (bottom). Boxplots of Gini coefficients of features are given specifically for color transforms: Red, Green, Blue, Saturation, Brightness, and OD Sum. For H&E, additional boxplots of the stain specific colors Hematoxilin, Eosin, and Residual are plotted. For ERG, additional boxplots of the stain specific color transforms Hematoxilin, ERG, and Residual are plotted. The prediction power of features varies for the color transforms in each staining.

In the following, we denote features with $p \leq 0.001$ as statistically significant. For H&E and ERG, we took the features of color transform Brightness that were significant. For PIN-4, we chose the significant features of color transform Saturation. The significant features were, ordered by decreasing Gini score:

- H&E staining, color transform Brightness, 12 significant features: F11, F12, F0, F4, F2, F9, F8, Median, Mean, F5, F7, F10.
- ERG staining, color transform Brightness, 13 significant features: Mean, F5, F7, F8, Median, F10, F4, F0, F9, F1, Std.dev., F3, F6
- PIN-4 staining, color transform Saturation, 16 significant features: Max, F12, F0, F7, F8, Median, Mean, F5, F9, F2, F4, F6, Std.dev., F3, F10, F1.

On these stain specific sets of features, we applied three ML algorithms: support vector machines classifier (SVM), neural networks (NN), and random forest (RF). Monte Carlo cross-validation with 100 random splits into 70% training set and 30% test determined the mean area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

Table 2 shows, in the rows denoted by ”default”, the mean AUC with standard deviation. The mean values of AUC are preferable high, $0.81 \leq \text{AUC} \leq 0.93$, and demonstrate the prediction power of the three groups of features for H&E, ERG, and PIN-4, respectively. The group of features of PIN-4 yields the best results. NN performs with the mean AUC of 0.93 ± 0.05 for the group of selected features of PIN-4. The range of values of AUC, however, do not exceed the values that might be expected from the Gini score of individual features. Note that, the Gini score of 0.865 of a single feature, Maximum Saturation of PIN-4, corresponds to an AUC of 0.93. Exemplary, Figure 2 shows the mean ROC curve of a Monte Carlo 100 random split cross-validation for NN and the 16 significant features in color transform Saturation of stain PIN-4.

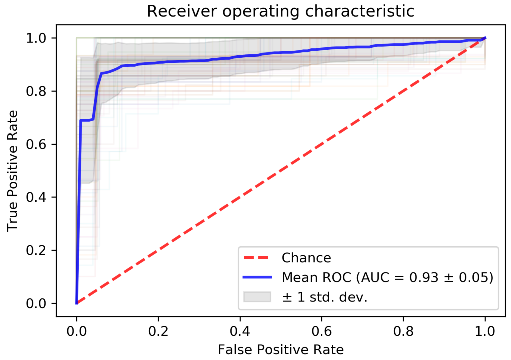

Figure 2: ROC curves of Monte Carlo cross-validation with 100 random splits. The blue curve denotes the mean ROC curve and the shaded gray area highlights its standard deviation. NN uses 12 significant features of the color transform Saturation of stain PIN-4 to compute the curves, see text. The mean value of AUC 0.93 ± 0.05 is preferable high and demonstrates the prediction power of the features of the color transform Saturation of stain PIN-4.

To enhance the performance of the algorithms, we applied an exhaustive grid search to optimize their hyper-parameters. For the values of the optimized hyper-parameters, we refer to Table 6 in Materials and Methods. Table 2 gives, in the rows denoted by ”tuned”, the mean AUC yielded by the classifiers with optimized parameters. The optimization of parameters improves the results for SVM and features of H&E, i.e., the mean AUC increases from 0.81 to 0.91. For all other combinations of ML algorithms and features sets, the optimization yields none or only minor improvement less than $\Delta\text{AUC} \leq 0.03$.

Since parameter optimizations yielded only small improvements of prediction power, we tested whether other groups of features may yield better results. We applied recursive feature elimination (RFE) [29] to select sets of non-redundant features from all color transforms of a stain. To reduce the computational expense of RFE, we removed shape features and redundant features by setting thresholds for the Pearson correlation coefficient. We chose thresholds for the Pearson correlation as large as possible but with the restriction to make the computation feasible of a conventional laptop computer with i7 processor and 8 GByte memory. The number of non-redundant features were $n = 84$ (PIN-4, Pearson correlation ≤ 0.95), $n = 104$ (H&E, Pearson correlation ≤ 0.99), and $n = 53$ (ERG, Pearson correlation ≤ 0.95). We applied RFE with successively increased numbers of top features and saved the set with highest accuracy. If two sets yielded identical accuracy, we favoured the smaller set. For the three stains, RFE yielded the sets:

- H&E staining, 25 features with mean accuracy 0.780 ± 0.061:
  Hematoxylin: Haralick Contrast (F1), Haralick Entropy (F8), Eosin: Haralick Contrast (F1), Haralick Sum of squares (F3), Haralick Sum variance (F6), Residual: Max, Haralick Correlation (F2), Haralick Sum variance (F6), Haralick Information measure of correlation 1 (F11), Haralick Information measure of correlation 2 (F12), Green: Haralick Contrast (F1), Haralick Sum average (F5), Haralick Sum entropy (F7), Haralick Difference entropy (F10), Blue: Haralick Contrast (F1), Haralick Sum entropy (F7), Haralick Entropy (F8), Haralick Difference entropy (F10), Brightness: Min, Saturation: Haralick Contrast (F1), Haralick Entropy (F8), Haralick Difference

Table 2: Mean AUC for three ML algorithms, support vector machines classifier (SVM), neural network (NN), and random forest (RF), trained on the sets of features from three stains, H&E (n=12), ERG (n=13), and PIN-4 (n=16), see text.

	SVM		RF		NN	
	default	tuned	default	tuned	default	tuned
H&E	0.81 ± 0.08	0.91 ± 0.05	0.82 ± 0.06	0.83 ± 0.07	0.85 ± 0.07	0.88 ± 0.07
ERG	0.83 ± 0.05	0.86 ± 0.06	0.85 ± 0.06	0.85 ± 0.06	0.86 ± 0.06	0.86 ± 0.06
PIN-4	0.92 ± 0.06	$\mathbf{0.94} \pm 0.05$	0.91 ± 0.05	0.92 ± 0.05	0.93 ± 0.05	$\mathbf{0.94} \pm 0.04$

entropy (F10), and OD Sum: Max, Haralick Entropy (F8), Haralick Difference entropy (F10).

- ERG staining, 9 features with mean accuracy 0.829 ± 0.066:
  Red: Median, Haralick Sum of squares (F3), Green: Mean, Haralick Contrast (F1), Brightness: Haralick Contrast (F1), Haralick Sum variance (F6), Haralick Sum of squares (F3), and OD Sum: Haralick Entropy (F8), Haralick Sum entropy (F7).
- PIN-4 staining, 5 features with mean accuracy 0.973 ± 0.037:
  Red: Median, Haralick Sum variance (F6), Blue: Haralick Sum of squares (F3), Haralick Contrast (F1), and Saturation: Haralick Sum variance (F6).

To compute reference accuracy values for the three sets of selected features, we applied NN and Monte Carlo cross-validation with 100 random splits into 70% training set and 30% test. With 25 selected features for H&E stain, NN discriminated with mean accuracy of 78.0% between malignant and benign cores. With nine selected features for ERG stain, NN reached a mean accuracy of 82.9%. With only five selected features for PIN-4, NN reached a favorable high mean accuracy of 97.3%.

Table 3 gives, the AUC for SVM, RF, and NN. The AUC scores are averaged over Monte Carlo cross-validation with 100 random splits. Compared to the corresponding AUC scores in Table 2, either no or minor improvement can be observed for stains H&E and ERG. For stain PIN-4, SVM and NN computed nearly perfect AUC of 0.99, see Figure 3 for the mean ROC curve of NN.

Discussion and conclusion

Automatically extracted features of the texture and intensity distribution of stained TMA turned out to be highly valuable to distinguish between malignant and benign tissue of the prostate gland. High prediction power could be shown already for individual features, as, e.g., Maximal Saturation of the PIN-4 stain. The three staining protocols H&E, ERG, and PIN-4 yielded different prediction power.

Table 3: Mean AUC values for three ML algorithms: SVM, RF, and NN, and three sets of features, H&E (n=25), ERG (n=9), and PIN-4 (n=5). The features are selected by recursive feature elimination (RFE), see text. Only for PIN-4 stain, RFE significantly improves the AUC compared to Table 2. With five features selected for stain PIN-4, SVM and NN reach a favorable high AUC$= 0.99 \pm 0.01$.

	SVM	RF	NN
H&E	0.83 ± 0.07	0.81 ± 0.07	0.82 ± 0.07
ERG	0.90 ± 0.05	0.87 ± 0.06	0.90 ± 0.05
PIN-4	$\mathbf{0.99} \pm \mathbf{0.01}$	0.95 ± 0.04	$\mathbf{0.99} \pm \mathbf{0.01}$

Cost-effective and simple H&E revealed promising results. Our results for H&E, e.g., AUC$= 0.91 \pm 0.05$,

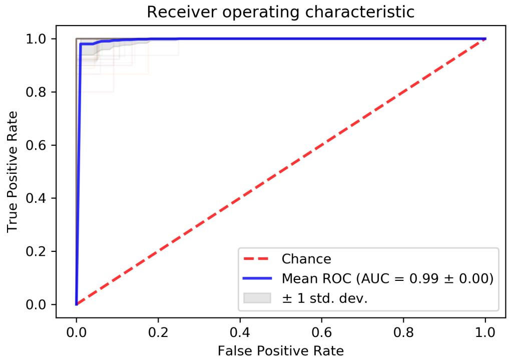

Figure 3: ROC curves of Monte Carlo cross-validation with 100 random splits. The blue curve denotes the mean ROC curve and the shaded gray area highlights its standard deviation. NN uses five features of the staining PIN-4 that are selected by recursive feature elimination (RFE), see text. The mean value of AUC 0.99 (mean accuracy 97.6%) demonstrates the power of the features of stain PIN-4 to discriminate between malignant and benign tissue cores.

SVM, tuned, were not impressive when compared to results of previous studies applying deep learning. Recently, AUC of in round numbers 0.99 have been reported for the application of convolutional neural networks to H&E stained TMA [30] and whole slide images [20]. Application of deep learning requires large number of data. Its black-box characteristic may be seen as possible drawback for medical decision making [31, 32, 30].

ERG stain revealed results of similar quality as H&E. Averaged over all cores, staining high expression of the proto-oncogene ERG seems to give no advantage compared to a simple H&E stain. In individual cases, the ERG stain may give valuable additional information. The size of our dataset did not enable us to test ERGs capability to potentially aid in differentiating variant states of prostate cancer aggressiveness. We abstained from testing the possible advantage of the combination of features of ERG stain with features of other stains, e.g., H&E and PIN-4.

In our study, PIN-4 showed the most accurate results. For PIN-4, SVM and NN yielded AUC$= 0.94 \pm 0.05$ for features extracted from color transform Saturation. PIN-4 has been reported to be useful in distinguishing prostatic adenocarcinoma from the benign mimickers [27, 33, 34]. We applied the stain PIN-4 as a cocktail of two antibodies, a brownish signal for high molecular weight cytokeratins, and a second, reddish signal for the protein alpha-methylacyl-CoA race-mase (AMCAR,P504S). P504S is a biomarker for prostate adenocarcinoma [35, 36]. Positive staining with a monoclonal antibody to high molecular weight cytokeratins has been shown to be of value in distinguishing between well-differentiated, small-acinar prostatic adenocarcinoma and its mimics [37]. Therefore, the superior performance of PIN-4 compared to H&E and ERG is not surprising. To our knowledge, PIN-4's potential application for automatic stratification of PCa in medical AI has not been tested up to now. In AI applications, staining with PIN-4 has been merely used as a preferable additional immunohistochemical workup to generate the ground truth by visual inspection [8, 22].

In the year 2010 Sabata *et al.* [27] studied the potential of computer aided diagnoses of PIN-4 stained needle biopsies. Their algorithm has identified the glands in the tissue and has classified the glands by the three simple criteria:

1 "If gland has only the brown basal staining then the tumor is benign"

2 "If gland has both the red racemace and the brown basal staining then it is classified as high-grade prostatic intraepithelial neoplasia (HG-PIN)"

3 "If gland has only the red racemace then it is classified as adenocarcinoma."

Sabata *et al.* have discussed several possible sources of missclassification. For small glands, a big variation in the intensity of racemace staining may cause recognition of the red staining to be error prone. It has been important to not merge a gland with the surrounding glands or the diagnosis would have been incorrect. Note that, automated object segmentation is a task and a potential source of missclassification. In view of recent studies, the three simple rules proposed by Sabata *et al.* are not likely to be able to account for possible heterogeneity of staining of benign and malignant tissue. It is possible that benign glands may show some weak to moderate AMCAR expression and on the other hand it is not a necessity for prostate cancer to be AMCAR positive (especially high grade subtypes can be negative and inter- and intratumoral heterogeneity can occur) [38, 39]. For benign tissue, e.g., tissue of atypical adenomatous hyperplasia (AHH), high expression of AMCAR has been reported in up to more than 50% of the cases [40].

In our approach, we used intensity and texture features of a core with, in general, multiple glands for its classification. Ranking features by their prediction power, i.e., their Gini coefficient, we found Maximum Saturation of PIN-4 by far the top feature. High Maximum Saturation indicated malignancy; 87% , i.e., 33 out of 38, malignant cores compared to only 6%, i.e., four out of 66, benign cores had Maximum Saturation above 0.953.

Why the presence of a pixel with high Saturation in a core was the best single indicator for malignancy can easily be understood. For rgb-values, Red, Green, Blue, of a pixel the Saturation, s, is determined by $s = 1 - \min(Red, Green, Blue)/\max(Red, Green, Blue)$. Left part of Figure 4 shows a malignant core (ID: RPX1:7B) with largest value of Maximum Saturation. Middle part of Figure 4 shows a region that contains a gland surrounded by basal cell with brownish membranous signal. The basal cells appear darker than cells in the left bottom part of the region with a pure reddish cytoplasmic signal. Right part of Figure 4 shows Saturation of the blow-up. The basal cells with brownish membranous signal (inside the outline of the gland) have a lower Saturation than cells with dominant reddish cytoplasmic signal for AMCAR (inside the circle, left bottom). The cells inside the outline vanished for the range of Saturation $0.91 \geq s \geq 1$.

Brownish membranous signal for cytokeratin yielded low values of Saturation whereas a pure reddish cytoplasmic signal for AMACR yielded high values of Saturation. Maximum Saturation identified a local region (2μm resolution) with reddish cytoplasmic signal and no brownish membranous signal for cytokeratin. The presence of such a local region inside a core with high

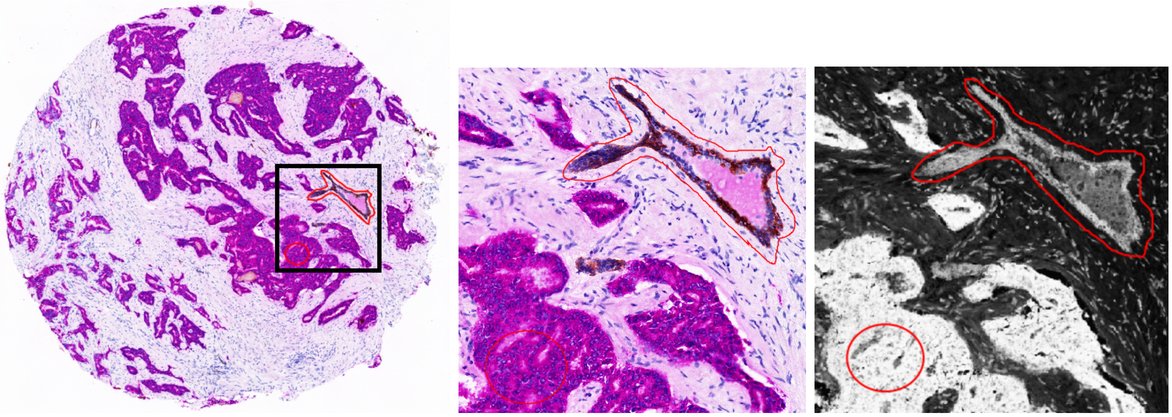

Figure 4: Left part: Malignant core (ID: RPX1:7B) stained with PIN-4. A zoom in region is marked by a rectangle. Middle part : Zoom into the marked region. In its upper right part, the blow–up shows brownish basal cells surrounding a gland and in the lower left part cells with dominant reddish cytoplasmic signal for AMACR. Right part: Saturation of the marked region. In the upper right part, a red outline indicate position and shape of the gland. In the left bottom part, a red circle indicate a region with dominant reddish signal. Brownish basal cells have lower Saturation than reddish cells.

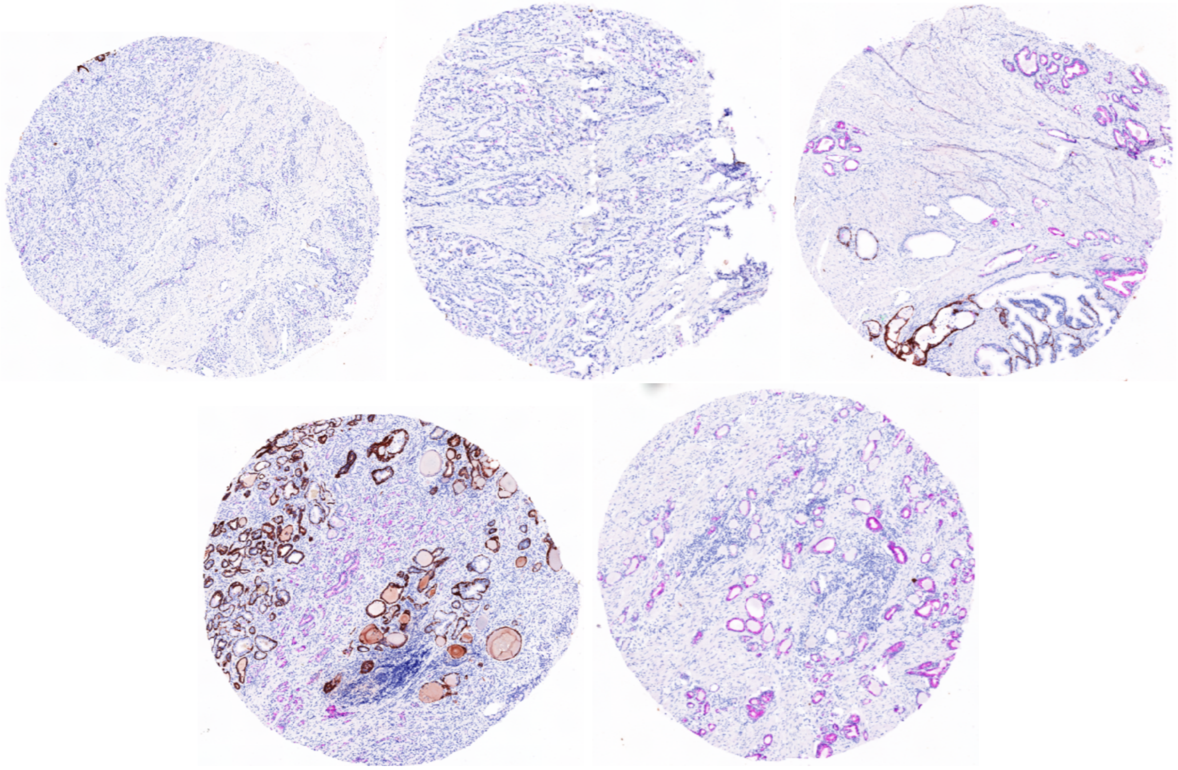

Figure 5: Malignant cores with Maximal Saturation below 0.953. Top left to bottom right, cores with ID/Maximum Saturation: RPX1:1A/0.808, RPX1:3C/0.889, RPX1:5E/0.894, RPX1:5B/0.919, and RPX1:7A/0.922. The cores represent high grade prostate cancer (two cores, top left) and low grade prostate cancers (remaining cores) with heterogenous AMCR expression.

and pure reddish signal was a good condition to diagnose malignancy. Our finding of 87% malignant cores and 6% benign cores with Maximum Saturation above 0.953 is in line with Murphy *et al.* [38] who described 91% and 11% AMCR positivity of prostate cancers and benign tissues, respectively.

Within our data set, we found five malignant cores with low values of Maximum Saturation. Figure 5 shows the five malignant cores with Maximum Saturation values below 0.953. Based on the threshold value 0.953 for Maximum Saturation, these five cores would had been missclassified as benign cores. The cores have low reddish signal if compared to cores with high Saturation as shown, e.g., in Figure 4. The low reddish signal for some malignant cores might be a results of the known potential inter- and intra-tumoral heterogeneity of prostate cancer AMACR positivity [38]. Heterogenous AMCAR expression is significantly associated with increased Gleason score and poorly differentiated tumors [38]. Respectively, the two tumor cores on the top left in Figure 5 show AMCAR negative high grade prostate cancers. The other tumor cores in Figure 5 show low grade prostate cancers with low AMCAR expression.

In our data set, we found also four benign cores with values of Maximum Saturation above 0.953, see Figure 6. Staining artefacts, see, e.g., core RPX3:1C, or intense dark brown staining, see, e.g., cores RPX3:2A, RPX3:7A, and RPX2:8C, lead to high values of Maximum Saturation. Three of these four benign cores were on the same slide, RPX3. Averaged exclusively over benign cores, Maximum Saturation 0.910 ± 0.043 of slide RPX3 was higher than Maximum Saturation 0.839 ± 0.029 and 0.873 ± 0.039 of slides RPX1 and RPX2, respectively. Compared to the mean value of Maximum Saturation 0.967 ± 0.037 of malignant cores, the mean values of benign cores on each slide were low. Slide to slide variations and the potentially heterogenous AMCAR expression of malignant and benign glands, however, made it problematic to determine a global threshold for a classification based solely on Maximum Saturation. Inclusion of definite benign tissue with and without AMCAR expression on a slide as reference for benign saturation values might be a potential solution.

Our patient cohort was biased towards older population (mean age, 66 ± 6.6 years). Since age-associated changes in AMACR expression has been reported in nonneoplastic prostatic tissues [41], age may be a valuable additional feature for populations with heterogeneous age distributions.

For PIN-4, the algorithms SVM and NN yielded even higher AUC of 0.99 ± 0.01 with five features that were extracted not only from Saturation but also from two additional color channels, Red, and Blue. The relevance of Red and Blue probably arose from the role of reddish signal for the protein alpha-methylacyl-CoA race-masse and the role of brownish signal for high molecular weight cytokerine, respectively. Saturation contributed to high value of AUC but, interestingly, not Maximum Saturation but Haralick Sum Variance of Saturation was one of the five selected features. In contrast to the local property Maximum Saturation, the Haralick Sum Variance measured a global property, i.e., a normalized value averaged over all neighbored pixel pairs of a core. In view of the slide to slide variation of staining in our data set, the application of a global and normalized Haralick feature may be more robust than the local measure of Maximum Saturation.

Notice that, our feature selection procedure may suffer from overfitting, for a discussion of a possible bias we refer to Demirciouglu (2021) [42]. The application of strategies to avoid overfitting in feature selection requires a larger data set. Our AUC values of classification without feature selection may be more reliable and relevant for applications to independent data sets. Despite the possible bias by overfitting, the sets of selected features on its own may be valuable for studies with independent data set. In future studies, it may be worthwhile to test deep-learning algorithms on PIN-4 images, to evaluate images of different image resolutions, to develop a suitable color model for PIN-4, and to study strategies to correct for slide to slide differences in staining, e.g., by a reference tissue cores or automated slide specific normalization techniques.

Material and Methods

Patient Cohort

Tissue/tumor samples and patient data were provided by the University Cancer Center Frankfurt (UCT). Written informed consent was obtained from all patients and the study was approved by the institutional Review Boards of the UCT and the Ethical Committee at the University Hospital Frankfurt (project-number: SUG-4-2018). The project expands on the results of Bernatz *et al.* [43] and in total 418 patients with confirmed PCa who were treated with radical prostatectomy (RPX) between 2014 and 2019 were screened for study inclusion [43]. In the current study, contrary to Bernatz *et al.* [43] patients with neoadjuvant therapy prior to RPX (n=6) were included and 1 PCa patient had to be excluded due to an insufficient amount of PCa-tissue leading to final study cohort of 38 PCa patients, see Bernatz *et al.* [43] for details. As negative control, 10 patients with benign prostatic hyperplasia (BPH) who were treated with Holmium laser enucleation of the prostate (HoLEP) were used. Inclusion criteria for the HoLEP cohort was (I) suffering from BPH

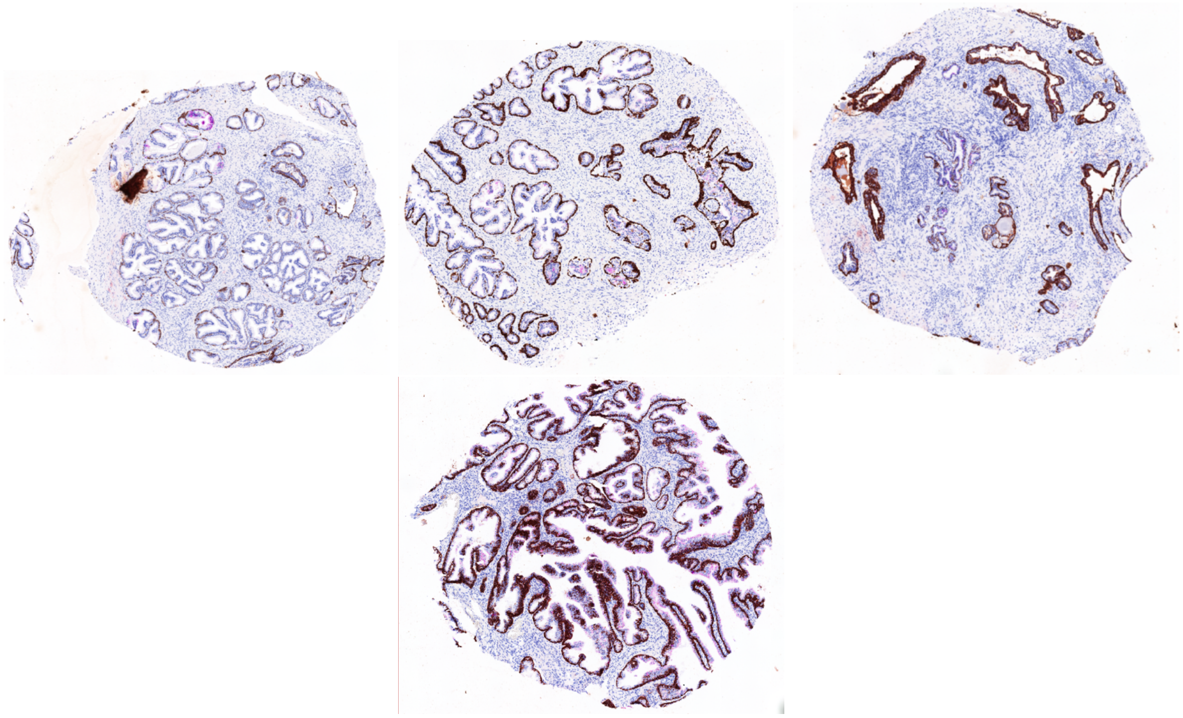

Figure 6: Benign cores with Maximum Saturation above 0.953. Top left to bottom right, cores with ID/Maximum Saturation: RPX3:1C/0.979, RPX3:2A/0.970, RPX3:7A/0.967, and RPX2:8C/0.964. Intense dark brown stain lead to high values of Maximum Saturation. For RPX3:1C (top left core), the staining artefact (dark spot in top left part of core) leads to its high value of Maximum Saturation.

and having received (II) treatment with HoLEP without (III) cancerous tissue in the HoLEP tissue. The final patient cohort comprised of 48 patients (mean age, 66 ± 6.6 years), 38 patients with PCa and ten patients with BPH.

Preparation of tissue microarrays

Prior to the TMA establishment, all whole slide specimen were annotated by an uropathologist (JK, 10 years of experience) to (I) delineate the areas of PCa index lesion with highest international society of urological pathology (ISUP) score, (II) benign tissue at the opposite site of the respective PCa slides, and (III) benign HoLEP specimen. In total, 48 paraffin-embedded tissue samples from our patient cohort were used to construct the TMAs by punching 106 representative tissue cores from the paraffin blocks. The representative punch-locations were annotated on respective H&E-slides of each tissue block which was used as a mask to identify respective regions on the tissue block. We punched a core (2mm diameter) from the index lesion of each PCa-tissue ($n = 38$). As matched-controls we used a tissue punch from the benign opposite site of each PCa whole gland specimen ($n = 38$) and three independent tissue-punches from each patient who was treated with HoLEP for benign prostatic hyperplasia ($n = 10 \times 3 = 30$). In total, 106 cores of prostate tissue (malignant, n=38; benign opposite site of PCa-patients, n=38; repetitive punches of HoLEP tissue, n=30) were punched to constuct TMA 1-3. TMA 1, TMA 2, and TMA 3 contained 42, 42, and 22 cores of prostate tissue, respectively, see Figure 7. TMA blocks were cut into 3μm thick slices and placed on an adhesive glass slide. Unstained slides were stained with H&E as well as with immunohistochemical staining ERG and PIN-4.

Histological staining

For immunohistochemistry (IHC), we used DAKO FLEX-Envision Kit (Agilent, Santa Clara, CA, US) and the fully automated DAKO Omnis staining system (Agilent, Santa Clara, CA, US) according to manufacturer´s instruction. We applied heat induced epitope retrieval at 97°C in high pH buffer, EnV FLEX TRS High pH Buffer (Agilent, Santa Clara, CA, US). Afterwards we applied immunohistochemical epitope staining for 20 min by either PIN-4 double stain or ERG single stain. PIN-4 co-stained high molecular weight cytokeratin, DAKO primary antibody Cytokeratin High

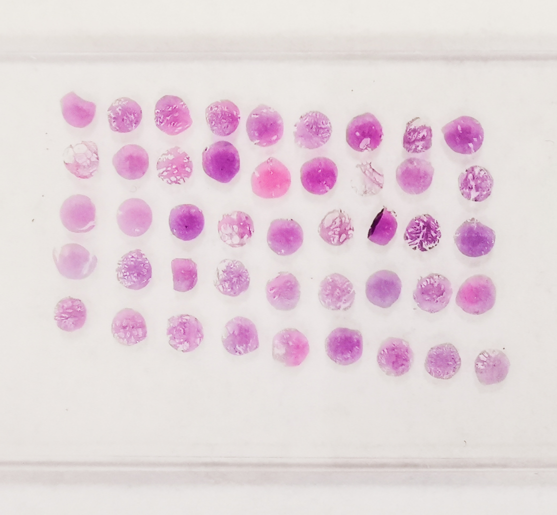

Figure 7: A TMA slide established at the Institute of Pathology, University Hospital Frankfurt. The tissue block with 45 cores, each 2 mm in diameter, is stained with H&E.

Molecular Weight (Clone 34betaE12, GA051, ready to use dilution, Agilent, Santa Clara, CA, US), and protein alpha-methylacyl-CoA racemase, AMACR (Clone 13H4, GA060; ready to use dilution, Agilent, Santa Clara, CA, US). ERG contained single-staining ERG primary antibody (GA659, Clone EP111, ready to use dilution, Agilent, Santa Clara, CA, US). For epitope visualization, we applied DAKO EnVision™ FLEX DAB+ and Magenta Substrate Chromogen System (Agilent, Santa Clara, CA, US). PIN-4 double stain produced a brownish membranous signal for cytokeratin and reddish cytoplasmic signal for AMACR. ERG single stain produced a brownish nuclear signal for high concentration of the protein ERG. After immunohistochemical staining, we used hematoxylin, DAKO hematoxylin solution (Agilent, Santa Clara, CA, US), for counterstaining. Hematoxylin produced blue-purple signal for cell nuclei.

For hematoxylin and eosin stain (H&E), slides were automatically processed using Tissue-Tek Prisma Plus staining device (Sakura Finetek) and Mayer´s Hematoxylin (AppliChem, Darmstadt, Germany) and Eosin (Waldeck, Münster, Germany) according to manufacturer´s instruction. H&E produced blue-purple signal for acidic cell nuclei and a pink signal for alkaline cytosolic and extracellular structures. Figure 8 shows three exemplary cores that are stained with H&E, ERG, and PIN-4, respectively.

Digitalization

We digitised the histologic slides with a digital slide scanner (Sysmex GmbH, Germany, resolution 2μm per pixel). We processed the images with an open source software for digital pathology and whole slide image analysis, QuPath (version 0.2.0) [28]. The image processing included de-arraying of the TMA and computation of feature values for each core.

Out of a total number of 318 stained cores, 106 cores times three stains, five cores had to be excluded from our analysis due to poor staining quality. The five excluded cores could not be recognized and processed by QuPath. For the detailed number of processed malignant and benign cores, we refer to Table 4.

Table 4: Number of malignant and benign cores stained with H&E, ERG, and PIN-4 and processed with the software QuPath.

staining	recognized cores	malignant cores	benign cores
H&E	105	36	69
ERG	104	35	69
PIN-4	104	36	68

QuPath extracted a grey-scale image for each color transform Red, Green, Blue (RGB color model), Hue, Saturation, Brightness (HSB color model) and Optical Density sum (OD–sum), see [44, 28] for a detailed description of the color models. We applied color deconvolution of QuPath to correct for minor variations between individual slides. For H&E, color deconvolution determined slide specific color vectors for Hematoxylin, Eosin, and Residual. For ERG, color deconvolution determined slide specific color vectors for Hematoxylin, ERG, and Residual. For PIN-4, color deconvolution determined slide specific color vectors that varied strongly from slide to slide and an unique assignment to stains was not possible. The limited ability of automated deconvolution to account for more than two stains may be the reason for the failure of color deconvolution in the case of the triple staining PIN-4.

We chose the set of standard features of QuPath. QuPath computed standard features for each core, as, e.g., five features of the intensity distribution, thirteen Haralick features based on the co-occurrence matrices for the texture, and shape values, as, e.g., area, circularity, solidity, max/min diameter of the core [28]. Within QuPath, the Haralick features are denoted by abbreviations F0–F12, for a list of abbreviations of features we refer to Table 5. After elimination of features with missing values or zero variance, we recorded 167, 166, and 172 features values for a core stained with H&E, ERG, and PIN-4, respectively.

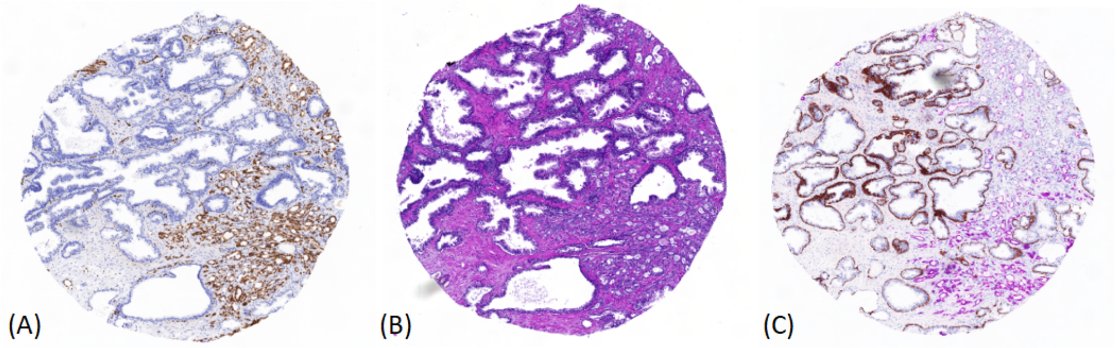

Figure 8: Three exemplary cores from TMAs demonstrating three stains: (A) H&E, (B) ERG, and (C) PIN-4. The images show subsections of TMA prepared at the Institute of Pathology, University Hospital Frankfurt.

Table 5: Abbreviation of predefined standard features in QuPath [28]. Five of the standard features are based on the intensity distribution and 13 are Haralick texture features.

Intensity-based basic features (5)	**Abbreviation**
Mean value	Mean
Standard deviation	Std
Minimum value	Min
Maximum value	Max
Median value	Median
Intensity-based Haralick features (13)	**Abbreviation**
Angular second moment	F0
Contrast	F1
Correlation	F2
Sum of squares	F3
Inverse difference moment	F4
Sum average	F5
Sum variance	F6
Sum entropy	F7
Entropy	F8
Difference variance	F9
Difference entropy	F10
Information measure of correlation 1	F11
Information measure of correlation 2	F12

Statistical Analysis

We applied the non-parametric Mann–Whitney U test [45] to compare two unpaired groups, e.g., malignant versus benign tissue. The Mann-Whitney test computes the U statistic of two samples. The U statistic determines the significance of the inequality of the two groups and the Gini coefficient. The Gini coefficient can be scaled to the area under the receiver operating characteristic (ROC) curve (AUC) [46]. The AUC represents the probability that a randomly chosen subject is correctly classified. An AUC of 0.5 corresponds to a random choice and an AUC of 1.0 corresponds to a perfect discrimination between the two groups [47, 46]. To correct the significance for multiple testing, we applied a Bonferroni adjustment and computed the false discovery rate (FDR) by the Benjamini-Hochberg procedure [48].

Parameter optimization

We optimized hyperparameters with the function `model_selection.GridSearchCV` of the scikit-learn library (version 0.22.1) [49] in Python. For SVM, we adjusted the regularization parameter, C, kernel coefficient, γ, and kernel. For RF, we adjusted number of trees in the forest, `n_estimators`, the maximum depth of the tree, `max_depth`, randomness of the bootstrapping of the samples, `random_state`, the minimum number of samples required to split an internal node, `min_samples_split`, and the minimum number of samples required at a leaf node, `min_samples_leaf`. For NN, we adjusted structure of a network, `hidden_layer_sizes`, activation function, `activation`, learning rate schedule for weight updates, `learning_rate`, solver for weight optimization, `solver`, regularization term, α, maximum number of iterations, `max_iter`, and random number generation for weights and bias initialization, `random_state`. For the customized hyper–parameters, we refer to Table 6.

Software

We processed tissue microarrays with the open source software for digital pathology and whole slide image analysis QuPath (version 0.2.0) [28]. We wrote Python scripts (Python version 3.7.6) [50]) in Jupyter Notebook [51]. We used modules from the scipy package

Table 6: Customized hyperparameters of the classifiers support vector machines classifier (SVM), neural networks (NN), and random forest (RF). We performed an exhaustive grid search to enhance the precision of the classifiers for each individual staining, H&E, ERG, and PIN-4.

	SVM	**RF**	**NN**

H&E	$C = 1000$,
	$\gamma = 0.00001$,
	kernel='linear',
	probability=True
		random_state= 1,
		max_depth= 15,
		n_estimators= 500,
		min_samples_split= 2,
		min_samples_leaf= 1
			hidden_layer_sizes= $(1, 100)$,
			learning_rate='constant',
			random_state= 1,
			solver='lbfgs'

ERG	$C = 0.1$,
	$\gamma = 0.005$,
	kernel='rbf',
	probability=True
		n_estimators= 100,
		max_depth= 25
			hidden_layer_sizes= $(1, 100)$,
			activation='identity',
			$\alpha = 0.0001$

PIN-4	$C = 100$,
	$\gamma = 0.0001$,
	kernel='linear',
	probability=True
		random_state=1,
		max_depth= 15,
		n_estimators= 500,
		min_samples_split= 2,
		min_samples_leaf= 1
			hidden_layer_sizes= $(1, 100)$,
			activation='logistic',
			max_iter= 1000,
			random_state= 1,
			learning_rate='constant',
			solver='lbfgs',
			$\alpha = 0.001$

(version 1.4.1) [52] for statistical calculations and applied ML algorithms from the scikit-learn library (version 0.22.1) [49].

Acknowledgements

T.J.V., P.J.W., and S.B. would like to thank the Frankfurt Cancer Institute/German Cancer Consortium for the support and funding as part of the Discovery & Development Project 2019...

Funding

This work was supported in part by the LOEWE Center Frankfurt Cancer Institute (FCI) funded by the Hessen StateMinistry for Higher Education, Research and the Arts (III L 5 - 519/03/03.001 - (0015))...

List of Abbreviations

AI artificial intelligence.
AUC area under the curve.
BPH benign prostatic hyperplasia.
ERG stain with a antibody for ERG protein (ERG for Erythroblast transformation-specific Related Gene).
FDR false discovery rate.
HGPIN High-Grade Prostatic Intraepithelial Neoplasia.
HoLEP Holmium Laser Enucleation of the Prostate.
H&E hematoxilin and eosin.
ISUP international society of urological pathology score.
ML machine learning.
NN neural networks.
PCa Prostate Cancer.
PIN-4 double stain with two antibodies, AMACR(P504S), and high molecular weight cytokeratin.
QuPath open source software for digital pathology and whole slide image analysis QuPath (version 0.2.0) [28].
RF random forest.
RFE recursive feature elimination.
ROC receiver operating characteristic.
RPX radical prostatectomy.
SVM support vector machines.
TMA tissue microarray.
UCT University Cancer Center Frankfurt.

Ethics approval and consent to participate

Written informed consent was obtained from all patients and the study was approved by the institutional Review Boards of the UCT and the Ethical Committee at the University Hospital Frankfurt (project-number: SUG-4-2018).

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Informatics, Institute of Bioinformatics, Ludwig-Maximilians-Universität München, Amalienstraße 17, 80333 München, Germany. [2]Molecular Bioinformatics Group, Institute of Computer Science, Faculty of Computer Science and Mathematics, Robert-Mayer-Straße 11–15, 60325 Frankfurt am Main, Germany. [3]Department of Diagnostic and Interventional Radiology, Goethe University Frankfurt am Main, University Hospital Frankfurt, 60590 Frankfurt am Main, Germany. [4]Dr. Senckenberg Institute for Pathology, Goethe University Frankfurt am Main, University Hospital Frankfurt, 60590 Frankfurt am Main, Germany. [5]Neurological Institute (Edinger Institute), University Hospital Frankfurt, 60590 Frankfurt am Main, Germany. [6] Department of Urology, Goethe University Frankfurt am Main, University Hospital Frankfurt, 60590 Frankfurt am Main, Germany. [7]Frankfurt Cancer Institute (FCI), University Hospital Frankfurt, 60590 Frankfurt am Main, Germany.

References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
2. Mottet, N., Cornford, P., van den Bergh, R., et al.: EAU–EANM–ESTRO–ESUR–ISUP–SIOG Guidelines on Prostate Cancer. European Association of Urology, Arnhem, The Netherlands, https://uroweb.org/wp-content/uploads/EAU-EANM-ESTRO_ESUR_ISUP_SIOG-Guidelines-on-Prostate-Cancer-2021.pdf (2021)
3. Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A.: The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. The American journal of surgical pathology **40**(2), 244–252 (2016)

4. Kann, B.H., Hosny, A., Aerts, H.J.: Artificial intelligence for clinical oncology. Cancer Cell **39**(7), 916–927 (2021)
5. Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I.: An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine **3**(1), 1–10 (2020)
6. Greer, M.D., Lay, N., Shih, J.H., Barrett, T., Bittencourt, L.K., Borofsky, S., Kabakus, I., Law, Y.M., Marko, J., Shebel, H., *et al.*: Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study. European Radiology **28**(10), 4407–4417 (2018)
7. Tătaru, O.S., Vartolomei, M.D., Rassweiler, J.J., Virgil, O., Lucarelli, G., Porpiglia, F., Amparore, D., Manfredi, M., Carrieri, G., Falagario, U., *et al.*: Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives. Diagnostics **11**(2), 354 (2021)
8. Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P.-H.C., Steiner, D.F., Manoj, N., Olson, N., Smith, J.L., Mohtashamian, A., *et al.*: Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. JAMA oncology **6**(9), 1372–1380 (2020)
9. Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., *et al.*: Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology **21**(2), 222–232 (2020)
10. Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports **6**(1), 1–11 (2016)
11. Lenain, R., Seneviratne, M.G., Bozkurt, S., Blayney, D.W., Brooks, J.D., Hernandez-Boussard, T.: Machine learning approaches for extracting stage from pathology reports in prostate cancer. Studies in health technology and informatics **264**, 1522 (2019)
12. Roffman, D.A., Hart, G.R., Leapman, M.S., Yu, J.B., Guo, F.L., Ali, I., Deng, J.: Development and validation of a multiparameterized artificial neural network for prostate cancer risk prediction and stratification. JCO clinical cancer informatics **2**, 1–10 (2018)
13. Lee, G., Veltri, R.W., Zhu, G., Ali, S., Epstein, J.I., Madabhushi, A.: Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. European urology focus **3**(4-5), 457–466 (2017)
14. Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J.D., Kapur, S., Reuter, V., Grady, L., Kanan, C., Klimstra, D.S., *et al.*: Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Modern Pathology **33**(10), 2058–2066 (2020)
15. Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., *et al.*: Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Medical image analysis **50**, 167–180 (2018)
16. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
17. Nir, G., Karimi, D., Goldenberg, S.L., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Thompson, D.J., *et al.*: Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. JAMA network open **2**(3), 190442–190442 (2019)
18. Nagpal, K., Foote, D., Liu, Y., Chen, P.-H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., *et al.*: Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. NPJ digital medicine **2**(1), 1–10 (2019)
19. Kwak, J.T., Hewitt, S.M.: Multiview boosting digital pathology analysis of prostate cancer. Computer methods and programs in biomedicine **142**, 91–99 (2017)
20. Pantanowitz, L., Quiroga-Garza, G.M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Shach, A.A., Shalev, V., Vecsler, M., *et al.*: An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. The Lancet Digital Health **2**(8), 407–416 (2020)
21. Bulten, W., Balkenhol, M., Belinga, J.-J.A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., *et al.*: Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. Modern Pathology **34**(3), 660–671 (2021)
22. Perincheri, S., Levi, A.W., Celli, R., Gershkovich, P., Rimm, D., Morrow, J.S., Rothrock, B., Raciti, P., Klimstra, D., Sinard, J.: An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Modern Pathology, 1–8 (2021)
23. Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R.: Hematoxylin and eosin staining of tissue and cell sections. Cold Spring Harbor Protocols **2008**(5), 4986 (2008)
24. Miettinen, M., Wang, Z.-F., Paetau, A., Tan, S.-H., Dobi, A., Srivastava, S., Sesterhenn, I.: Erg transcription factor as an immunohistochemical marker for vascular endothelial tumors and prostatic carcinoma. The American journal of surgical pathology **35**(3), 432 (2011)
25. Adamo, P., Ladomery, M.: The oncogene erg: a key factor in prostate cancer. Oncogene **35**(4), 403–414 (2016)
26. Humphrey, P.: Diagnosis of adenocarcinoma in prostate needle biopsy tissue. Journal of clinical pathology **60**(1), 35–42 (2007)
27. Sabata, B., Babenko, B., Monroe, R., Srinivas, C.: Automated analysis of pin-4 stained prostate needle biopsies. In: International Workshop on Prostate Cancer Imaging, pp. 89–100 (2010). Springer
28. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W.: QuPath: Open source software for digital pathology image analysis. Scientific Reports **7**(1), 1–7 (2017)
29. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1), 389–422 (2002)
30. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G.: Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology **21**(2), 233–241 (2020)
31. Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., Stumpe, M.C.: Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. Archives of pathology & laboratory medicine **143**(7), 859–868 (2019)
32. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)
33. Dabir, P.D., Ottosen, P., Høyer, S., Hamilton-Dutoit, S.: Comparative analysis of three-and two-antibody cocktails to AMACR and basal cell markers for the immunohistochemical diagnosis of prostate carcinoma. Diagnostic pathology **7**(1), 1–6 (2012)
34. Chougani, S., Sunandalakshmi, G., Kharidehal, D., Ravisankar, V., Vissa, S.: Utility of PIN4 cocktail antibody in the atypical foci of the prostate. International Journal of Clinical and Diagnostic Pathology **3**(1), 396–403 (2020)
35. Xu, J., Stolk, J.A., Zhang, X., Silva, S.J., Houghton, R.L., Matsumura, M., Vedvick, T.S., Leslie, K.B., Badaro, R., Reed, S.G.: Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. Cancer research **60**(6), 1677–1682 (2000)
36. Jiang, Z., Woda, B.A., Rock, K.L., Xu, Y., Savas, L., Khan, A., Pihan, G., Cai, F., Babcook, J.S., Rathanaswami, P., *et al.*: P504s: a new molecular marker for the detection of prostate carcinoma. The American journal of surgical pathology **25**(11), 1397–1404 (2001)
37. O'Malley, F., Grignon, D., Shum, D.: Usefulness of immunoperoxidase staining with high-molecular-weight cytokeratin in the differential diagnosis of small-acinar lesions of the prostate gland. Virchows Archiv A **417**(3), 191–196 (1990)

38. Murphy, A., Hughes, C., Lannigan, G., Sheils, O., O'Leary, J., Loftus, B.: Heterogeneous expression of α-methylacyl-CoA racemase in prostatic cancer correlates with Gleason score. Histopathology **50**(2), 243–251 (2007)
39. Hasan, I.A., Gaidan, H.A., Al-Kaabi, M.M.: Diagnostic value of cytokeratin 34 beta E12 (Ck34βE12) and α-Methylacyl-CoA racemase (AMACR) immunohistochemical expression in prostatic lesions. Iranian Journal of Pathology **15**(3), 232 (2020)
40. Zhang, C., Montironi, R., MacLennan, G.T., Lopez-Beltran, A., Li, Y., Tan, P.-H., Wang, M., Zhang, S., Iczkowski, K.A., Cheng, L.: Is atypical adenomatous hyperplasia of the prostate a precursor lesion? The Prostate **71**(16), 1746–1751 (2011)
41. Gologan, A., Bastacky, S., McHale, T., Yu, J., Cai, C., Monzon-Bordonaba, F., Dhir, R.: Age-associated changes in alpha-methyl CoA racemase (AMACR) expression in nonneoplastic prostatic tissues. The American journal of surgical pathology **29**(11), 1435–1441 (2005)
42. Demircioğlu, A.: Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights into Imaging **12**(1), 172 (2021)
43. Bernatz, S., Ackermann, J., Mandel, P., Kaltenbach, B., Zhdanovich, Y., Harter, P.N., Döring, C., Hammerstingl, R., Bodelle, B., Smith, K., Bucher, A., Albrecht, M., Rosbach, N., Basten, L., Yel, I., Wenzel, M., Bankov, K., Koch, I., Chun, F.K.-H., Köllermann, J., Wild, P.J., Vogl, T.J.: Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features. European Radiology, 1–13 (2020)
44. Ruifrok, A.C., Johnston, D.A., *et al.*: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology **23**(4), 291–299 (2001)
45. Cheung, Y.K., Klotz, J.H.: The Mann Whitney Wilcoxon distribution using linked lists. Statistica Sinica, 805–813 (1997)
46. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of mathematical psychology **12**(4), 387–415 (1975)
47. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)
48. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 1165–1188 (2001)
49. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research **12**, 2825–2830 (2011)
50. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)
51. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C.e.a.: Jupyter Notebooks-a publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas, Proceedings of the 20th Confernce on Electronic Publishing, pp. 87–90 (2016)
52. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J.e.a.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods **17**(3), 261–272 (2020)