



OPEN Integrating radiological and clinical data for clinically significant prostate cancer detection with machine learning techniques

Luis Mariano Esteban^{1,2}✉, Ángel Borque-Fernando^{3,4,5}, Maria Etelvina Escorihuela¹, Javier Esteban-Escañó⁶, Jose María Abascal⁷, Pol Servian⁸ & Juan Morote^{9,10,11}

In prostate cancer (PCa), risk calculators have been proposed, relying on clinical parameters and magnetic resonance imaging (MRI) enable early prediction of clinically significant cancer (CsPCa). The prostate imaging–reporting and data system (PI-RADS) is combined with clinical variables predominantly based on logistic regression models. This study explores modeling using regularization techniques such as ridge regression, LASSO, elastic net, classification tree, tree ensemble models like random forest or XGBoost, and neural networks to predict CsPCa in a dataset of 4799 patients in Catalonia (Spain). An 80–20% split was employed for training and validation. We used predictor variables such as age, prostate-specific antigen (PSA), prostate volume, PSA density (PSAD), digital rectal exam (DRE) findings, family history of PCa, a previous negative biopsy, and PI-RADS categories. When considering a sensitivity of 0.9, in the validation set, the XGBoost model outperforms others with a specificity of 0.640, followed closely by random forest (0.638), neural network (0.634), and logistic regression (0.620). In terms of clinical utility, for a 10% missclassification of CsPCa, XGBoost can avoid 41.77% of unnecessary biopsies, followed closely by random forest (41.67%) and neural networks (41.46%), while logistic regression has a lower rate of 40.62%. Using SHAP values for model explainability, PI-RADS emerges as the most influential risk factor, particularly for individuals with PI-RADS 4 and 5. Additionally, a positive digital rectal examination (DRE) or family history of prostate cancer proves highly influential for certain individuals, while a previous negative biopsy serves as a protective factor for others.

Keywords Clinically significant prostate cancer, Machine learning, Clinical utility, SHAP values

Artificial intelligence (AI), particularly within the domain of machine learning (ML), has demonstrated notable advancements in the fields of medicine and biomedical research¹. Through the utilization of mathematical methodologies, AI facilitates the processes of decision-making, reasoning, and adaptation in intricate scenarios².

In 2020, prostate cancer (PCa) ranked as the second most frequently diagnosed cancer and the fifth leading cause of cancer-related mortality among men globally³. In 2014, the ISUP Gleason Grading Conference introduced grade groups (1–5) to enhance patient stratification⁴. ISUP grade 1 (Gleason 6) indicates indolent PCa (iPCa), often suitable for active surveillance, while ISUP > 1 (Gleason ≥ 7) identifies clinically significant PCa (csPCa), requiring treatment. The European Association of Urology (EAU) recommends a risk-stratified approach to prostate cancer (PCa) screening, emphasizing the use of Prostate-Specific Antigen (PSA) testing

¹Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, C/ Mayor 5, 50100 La Almunia de Doña Godina, Spain. ²Institute for Biocomputation and Physics of Complex Systems (BIFI), 50009 Zaragoza, Spain. ³Department of Urology, Miguel Servet University Hospital, 50009 Zaragoza, Spain. ⁴Area of Urology, Department of Surgery, Faculty of Medicine, University of Zaragoza, 50009 Zaragoza, Spain. ⁵Health Research Institute of Aragon Foundation, 50009 Zaragoza, Spain. ⁶Department of Electronic Engineering and Communications, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, 50100 La Almunia de Doña Godina, Spain. ⁷Department of Urology, Department of Surgery, Parc de Salut Mar, Universitat Pompeu Fabra, 08003 Barcelona, Spain. ⁸Department of Urology, Hospital Germans Trias i Pujol, 08916 Badalona, Spain. ⁹Department of Urology, Vall d'Hebron Hospital, 08035 Barcelona, Spain. ¹⁰Department of Surgery, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. ¹¹Research Group in Urology, Vall d'Hebron Research Institute, 08035 Barcelona, Spain. ✉email: lmeeste@unizar.es

as a primary tool in early detection. The goal is to identify clinically significant prostate cancer (csPCa) while minimizing the risks of overdiagnosis and overtreatment, common issues associated with PSA-based screening⁵. The EAU guidelines suggest that the use of PSA should be integrated with other diagnostic measures, such as digital rectal exams (DRE) and increasingly, magnetic resonance imaging (MRI), to improve the detection of csPCa⁶.

In the Prostate MRI imaging Study (PROMIS)⁷, MP-MRI demonstrated significantly higher sensitivity (93%) compared to transrectal ultrasound-guided biopsy (TRUS, 48%) for detecting clinically significant prostate cancer (csPCa) while reducing unnecessary biopsies by 27%. Using MP-MRI as a triage test prior to biopsy improved diagnostic accuracy and reduced overdiagnosis of clinically insignificant cancer. PI-RADS v2.1 is widely adopted for evaluating suspicious lesions, with thresholds (≥ 3) indicating biopsy necessity based on a semiquantitative risk scoring system. Negative predictive value ranges between 96 and 98% for PI-RADS 1 and 2 lesions, confirming its utility in reducing unnecessary biopsies. Despite significant improvements, challenges remain in inter-reader variability and overdiagnosis of indolent prostate cancer, particularly in PI-RADS 3 lesions. Radiomics, when integrated with PI-RADS, shows promise in enhancing predictive accuracy for csPCa by leveraging AI-driven image feature analysis⁸.

Contemporary AI predominantly relies on ML techniques, adept at discerning patterns in vast datasets and constructing reasoning systems for patient risk assessment and improved decision-making⁹. Data mining methods and adaptive ML algorithms consistently surpass traditional statistical approaches in facilitating more precise patient-level predictions and modeling disease prognosis and risk. ML-based techniques offer the advantage of automating hypothesis formulation and evaluation, while also assigning parameter weights to predictors based on their correlation with outcome predictions¹⁰. However, the considerable potential of AI in cancer research must be cautiously considered in light of the challenges concerning transparency and reproducibility that need to be addressed¹¹.

MRI-based predictive models personalize the probability of CsPCa and enhance the selection of candidates for prostate biopsy. Risk calculators (RCs) are available for this purpose, although many are solely based on logistic regression models¹². The design of web applications has facilitated the widespread adoption of their usage. The Rotterdam (ROT) MRI-RC was developed adapting the ROT-RC to data from five centers from Germany and Netherlands in biopsy-naïve men and those with previous negative prostate biopsy, in 961 men with serum PSA > 3.0 ng/ml and/or abnormal digital rectal exam (DRE) findings¹³. Recently, the Barcelona (BCN) MRI-RC was developed among 1,486 men using data with the same characteristics¹⁴. Both demonstrated robust external validation in a cohort comprising 946 men, as indicated by area under the ROC curve values of 0.856 and 0.844 for BCN-RC and ROT-RC, respectively. Furthermore, they exhibit notable clinical utility, selecting a cutoff point corresponding to a 10% misclassification rate of CsPCa, resulting in a reduction of biopsy procedures by 35 and 36% for BCN-RC and ROT-RC, respectively¹⁵.

Recent advancements in integrating magnetic resonance imaging (MRI) features with artificial intelligence (AI) models have shown promise in enhancing the detection and characterization of clinically significant prostate cancer (csPCa). However, there is a need for clearer delineation of methodologies and incorporation of complementary clinical parameters, such as prostate-specific antigen (PSA) density, to improve model robustness and clinical applicability.

MRI-based imaging features, such as those derived from the prostate imaging-reporting and data system (PI-RADS) scoring and radiomic analyses, have become pivotal in AI-driven csPCa detection models. Handcrafted radiomics extract predefined quantitative features from T2-weighted, ADC, and DWI sequences, while deep learning methods automate the extraction of complex, high-dimensional features.

The primary objective of this study is to analyze the predictive capability of various machine learning models in forecasting clinically significant prostate cancer by combining the PI-RADS score with clinical variables. Specifically, emphasis is placed on assessing the clinical utility of the models and their explicability, with attention directed towards the role of each variable in the prediction process.

Materials and methods

Data recruitment

A retrospective analysis was carried out involving 5005 men with serum PSA levels of ≥ 3.0 ng/ml and/or an abnormal DRE. These men were recruited from 10 centers of the Catalan CsPCa early detection program between 2018 and 2022. All included men underwent prebiopsy mpMRI and targeted and/or systematic biopsies. Men who received 5-alpha reductase inhibitors (5-ARI) and had a history of previous PCa, atypical small acinar proliferation, or high-grade prostatic intraepithelial neoplasia with atypia were excluded from the study. After applying the exclusion criteria, we found 4799 participants. The project received ethical approval from the VHH ethical committee (PR/AG-317/2017). Written informed consent was obtained from all subjects involved in the study. This study was carried out following the ethical guidelines of the Declaration of Helsinki 1964 (version 2000) and the legislation involving humans.

The objective of the study was to anticipate CsPCa, which was defined as an International Society of Urologic Pathology grade group of ≥ 2 ⁴. To achieve this objective, the study gathered patient data, encompassing multiple variables such as Age, PSA values, prostate volume, PSA density, digital rectal examination (DRE) findings, family history of prostate cancer, history of previous negative biopsies, and PI-RADS scores. PSAD (ng/mL²) was calculated from the pre-biopsy serum PSA and the prostate volume reported in the pre-biopsy MRI.

Statistical analysis

A descriptive analysis of the data was conducted to compare patients with and without CsPCa. Continuous variables were summarized using the median and interquartile range (IQR), while categorical variables were summarized using absolute and relative frequencies for each category. To assess differences between the groups

with CsPCa and non-CsPCa conditions, the Mann-Whitney test or Chi-square test was employed for continuous and categorical data, respectively. Additionally, a description of the variables under study was conducted for each center.

Machine learning models

To predict CsPCa, various machine learning models were developed, including logistic regression, Ridge regression, LASSO, elastic net, classification tree, random forest, XGBoost and neural network. The original database was randomly divided into training (70%) and validation (30%) datasets to develop and evaluate these models. Care was taken to ensure that both groups had a similar proportion of CsPCa cases.

Logistic regression

The construction of the logistic regression model involved a stepwise selection process, employing a backward/forward method. This iterative process involved removing variables based on an improvement in the Akaike index, while also considering the inclusion of variables that were removed from the model if their inclusion improved the index at any step.

Ridge regression, LASSO and elastic net

Ridge regression, LASSO (least absolute shrinkage and selection operator), and elastic net are regularization techniques employed to prevent multicollinearity and reduce dimensionality¹⁶.

Ridge Regression introduces a penalty term to the objective function, multiplying the squares of the coefficients multiplied by a regularization parameter (λ), facilitates a shrinkage effect on the coefficients towards zero without exacting them to zero. LASSO, instead of the sum of squared coefficients, employs the sum of the absolute values of coefficients multiplied by the regularization parameter. LASSO exhibits a distinct characteristic of effecting variable selection by zeroing out certain coefficients. Elastic Net amalgamates the penalties from ridge regression and LASSO, integrating both L1 (LASSO) and L2 (ridge regression) penalties into the linear regression objective function, each governed by distinct regularization parameters. The optimization of λ was performed using AUC as the objective parameter.

Classification tree

Classification trees are recursive partition models that minimize the impurity of the classes defined by the partition¹⁷. They provide a simple classification system that is easy to implement, but they often lack high discrimination ability. In this study, we used the Gini index as the loss function and set the minimum number of observations required for a split to 50 in a node. Additionally, we set the minimum number of observations in any terminal node to 17 and limited the maximum depth of any node in the final tree to 30.

Random forest

Random forests consist of an ensemble of classification trees, where each tree is trained using a unique bootstrap sample and a different combination of variables¹⁸. This approach renders a bagging classifier that ensures diversity among the trees, resulting in a more robust model. For our analysis, we utilized a sample size of 2426, which corresponds to approximately 60% of the total data, to build each tree. The splitting rule employed was Gini, and we considered 10 random split points. The error rate was evaluated for the CsPCa cases (1), non-PCa cases (0), and all data, using the out-of-bag (OOB) error estimation. The data not used for building each tree was used to estimate prediction error, and an average prediction error was estimated. To optimize the terminal node size of the forest and the number of variables used to train each tree, we employed a tuning optimization parameter.

Extreme gradient boosting

Extreme Gradient Boosting (XGBoost) belongs to the ensemble learning family and sequentially builds a series of weak learners (decision trees) to create a strong predictive model¹⁹. XGBoost incorporates regularization techniques, handles missing values, and employs parallel and distributed computing to accelerate training. It is widely used in various domains due to its performance, flexibility, and feature importance analysis.

Neural networks

Neural networks provide a wide range of non-linear prediction models. For this analysis, we employed a classical perceptron with a single hidden layer. The neural network was trained with different architectures, using different random initial weights and training parameters. We experimented with learning rates of 0.05 and 0.1, and used the hyperbolic tangent (tanh) and logistic activation functions.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}; \text{ logistic}(x) = \frac{1}{1 + e^{-x}}$$

Model interpretability

Furthermore, for the purpose of studying the explainability of machine learning models, we conducted an analysis of variable importance using SHAP (SHapley Additive exPlanations) values. Shapley values were computed to determine the significance of each variable in individual predictions, while summary plots collectively provide insights into the importance of variables across the entire cohort's predictions. These values are derived through the application of game theory, which examines all feasible combinations of features and their effects on prediction outcomes. The computation of the Shapley value involves systematically perturbing input features and observing how alterations to these features correlate with the resultant model predictions. Subsequently, the Shapley value for a specific feature is determined as the average marginal contribution it makes to the overall

model score. This methodology offers a rigorous means to discern the individual influence of features on model predictions, providing valuable insights into their respective contributions to the model's output²⁰.

Model validation

For validation purposes, the models underwent evaluation in terms of discrimination and clinical utility. Discrimination was evaluated using the receiver operating characteristic (ROC) curve. Choosing a specific cutoff point for the CsPCa probability, individuals will be classified as CsPCa or non-CsPCa based on whether the probability assigned by the model is above or below the established threshold. Since our model is not perfect, we will correctly classify some patients as CsPCa (true positives, TP) or non-CsPCa (true negatives, TN) at times, but there will also be misclassified cases, including both non-CsPCa (false negatives, FN) and CsPCa (false positives, FP) instances. The ROC curve illustrates pairs of Sensitivity (True positive rate (TP / (TP + FN)), y-axis) versus 1-Specificity (False positive rate (FN / (TN + FP)), X-axis across different CsPCa probability cutoff values. The area under the ROC curve (AUC) summarizes the predictive model's discrimination ability. AUC measures the likelihood that the model assigns a higher probability of CsPCa to an actual CsPCa case compared to a non-CsPCa case. It ranges from 0 to 1, where 0.5 signifies randomness, 0.7 is considered acceptable, 0.8 suggests strong performance, 0.9 indicates excellent performance, and 1 signifies perfect discrimination. The AUC's 95% confidence intervals were computed using DeLong estimation, also AUC comparison was based on DeLong test²¹. Moreover, specificities were investigated and compared for various sensitivity thresholds (0.8, 0.85, 0.9, 0.95) using a proportion test.

Furthermore, the practical applicability of the developed machine learning models was assessed by evaluating their clinical utility. This evaluation involved treating the prediction models as dichotomous classification models, utilizing a specific cutoff point to differentiate between positive (CsPCa) and negative (non-CsPCa) individuals above or below the threshold. To assess this issue, the clinical utility curve²² was employed. This curve deployed the threshold probability on the X-axis to identify patients as CsPCa cases, while the Y-axis indicates the percentage of two distinct measures. The first measure represents the percentage of CsPCa cases incorrectly classified below the chosen cutoff point, while the second measure indicates the number of patients falling below the cutoff point. Analyzing this curve for different cutoff points enables determination of the percentage of misclassified CsPCa and patients with very low risk of CsPCa who could avoid unnecessary biopsies. These parameters hold significant importance in clinical practice.

Software

The statistical analyses were executed using the R programming language version 4.2.2 (The R Foundation for Statistical Computing, Vienna, Austria). Several libraries were employed, such as `regplot`, `rpart`, `randomForestSRC`, `xgboost`, `SHAPforxgboost`, `nnet` `NeuralNetTools` `shapviz` and `kernelshap`²³. Validation procedures utilized the `PROC` R library, along with the `CUC` R code function.

Results

Descriptive characteristics

The cohort of the study comprises data from 10 different centers, we found variability within the database. Table 1 shows the characteristics of study variables in each centre. CsPCa ranges from 35 to 65% render heterogeneity in our database. Also, we found differences in predictor variables.

To perform our comparative machine learning analysis, data was split in training and validation cohort. Table 2 provides an overview of the patients characteristics. We observed similar behaviour between development and validation cohorts. With the exception of familiar antecedents of prostate cancer in training dataset, all variables show significant differences between CsPCa and non-CsPCa groups.

Multivariate prediction models

In order to predict CsPCa, we employed both conventional and machine learning approaches for classification problems. The conventional approach involved utilizing the stepwise logistic regression model. Additionally, we utilized regularization techniques in ridge regression, LASSO and elastic net models, and machine learning algorithms such as, classification trees, random forest, XGBoost and artificial neural networks. Models were built using training data, and their discrimination, calibration, and clinical utility were estimated using validation data.

Logistic regression

Table 3 displays the variables that were found to be statistically significant in the multivariate analysis. Regarding the PI-RADS variable, PI-RADS 1 was used as the reference category. In comparison to this category, PI-RADS 2 is not significant, while categories 3, 4, and 5 are significant, demonstrating a higher risk of clinically significant prostate cancer (CsPCa) than category 1.

To illustrate the weight of the variables in the prediction model we provide a nomogram in (Fig. 1). The nomogram shows the weight of variables in the predicted probability of CsPCa. For each individual, a score is assigned to each variable on the upper axis. By summing up these scores, a total score is obtained, which provides us with the probability of CsPCa on the lower axis. Considering the variability of points assigned on the nomogram, the variables that show the most predictive ability were the score assigned by PSA, prostate volume and the PI-RADS score categories.

Ridge regression, LASSO and elastic net

Table 4 shows the coefficients of the models corresponding to the best parameter λ , that were 0.02244428, 0.0007699442 and 0.0005743105 for Ridge regression, LASSO and Elastic net respectively. In the case of the

Variable	Center									
	1	2	3	4	5	6	7	8	9	10
N	70	74	25	178	280	588	1116	243	196	2030
CsPCa	27 (39%)	31 (42%)	14 (56%)	62 (35%)	128 (46%)	213 (36%)	541 (48%)	87 (36%)	127 (65%)	772 (38%)
Age	68 (63,73)	66 (59,73)	67 (58,72)	69 (64,73)	69 (62,74)	67 (61,72)	67 (62,73)	67 (63,74)	66 (61,73)	69 (62,74)
PSA	6.3 (5.0,9.1)	7.6 (5.7,10.8)	9.0 (7.0,10.6)	7.7 (5.6,12.0)	9.9 (6.3,20.9)	7.6 (5.8,10.2)	7.5 (5.6,12.0)	6.2 (4.8,8.4)	6.8 (4.8,9.6)	6.1 (4.4,9.8)
PV	61 (44,80)	55 (38,73)	50 (39,62)	54 (37,80)	55 (41,76)	60 (42,84)	52 (38,74)	58 (42,72)	45 (32,64)	54 (40,76)
PSAD	0.13 (0.08,0.17)	0.15 (0.10,0.26)	0.18 (0.11,0.25)	0.14 (0.09,0.24)	0.19 (0.11,0.38)	0.13 (0.09,0.20)	0.14 (0.09,0.25)	0.11 (0.08,0.16)	0.15 (0.10,0.21)	0.12 (0.07,0.19)
DRE	9 (13%)	15 (20%)	9 (36%)	75 (42%)	52 (19%)	195 (33%)	413 (37%)	45 (19%)	37 (19%)	502 (25%)
FH	14 (20%)	13 (18%)	1 (4%)	9 (5%)	16 (6%)	51 (9%)	38 (3%)	4 (2%)	33 (17%)	162 (8%)
PNB	65 (93%)	20 (27%)	8 (32%)	79 (44%)	154 (55%)	236 (40%)	222 (20%)	102 (42%)	58 (30%)	508 (25%)
PI-RADS										
1	0 (0%)	9 (12%)	0 (0%)	53 (30%)	16 (6%)	4 (1%)	242 (22%)	0 (0%)	0 (0%)	225 (11%)
2	0 (0%)	5 (7%)	0 (0%)	5 (3%)	2 (1%)	18 (3%)	63 (6%)	0 (0%)	1 (1%)	74 (4%)
3	2 (3%)	10 (14%)	1 (4%)	22 (12%)	59 (21%)	154 (26%)	221 (20%)	40 (16%)	31 (16%)	667 (33%)
4	50 (71%)	38 (51%)	13 (52%)	57 (32%)	121 (43%)	310 (53%)	397 (35%)	170 (70%)	11 (56%)	659 (32%)
5	18 (26%)	12 (16%)	11 (44%)	41 (23%)	82 (29%)	102 (17%)	193 (17%)	33 (14%)	53 (27%)	405 (20%)

Table 1. Characteristics of study variables by center. *CsPCa* clinically significant prostate cancer, *PSA* prostate specific antigen, *PV* prostate volume, *PSAD* PSA density, *DRE* digital rectal exam findings. *FH* family history of PCa, *PNB* previous negative biopsy.

Variable	Development cohort <i>n</i> = 3839			Validation cohort <i>n</i> = 960		
	CsPCa <i>n</i> = 1607 (41.9%)	Non-CsPCa <i>n</i> = 2232 (58.1%)	p-value	CsPCa <i>n</i> = 394 (41.1%)	Non-CsPCa <i>n</i> = 566 (58.9%)	p-value
Age	71 (65,76)	66 (60,71)	<0.001	70 (65,77)	66 (61,72)	<0.001
PSA	8.0 (5.6,13.5)	6.3 (4.7,9.0)	<0.001	8.7 (5.7,14.9)	6.2 (4.5,9.0)	<0.001
PV	45 (34, 62)	62 (44,85)	<0.001	47 (35,65)	64 (46,86)	<0.001
PSAD	0.18 (0.12,0.32)	0.10 (0.07,0.16)	<0.001	0.19 (0.12,0.33)	0.10 (0.07,0.15)	<0.001
DRE (suspicious)	699 (43.5%)	368 (16.5%)	<0.001	186 (47.2%)	99 (17.5%)	<0.001
FH (yes)	125 (7.8%)	148 (6.6%)	0.172	36 (9.1%)	32 (5.7%)	0.038
PNB (yes)	392 (24.4%)	768 (34.4%)	<0.001	87 (22.1%)	205 (36.2%)	<0.001
PI-RADS			<0.001			<0.001
1	51(3.2%)	393 (17.6%)		9 (2.3%)	96 (17.0%)	
2	21 (1.3%)	117 (5.2%)		2 (0.5%)	28 (4.9%)	
3	163 (10.1%)	801 (35.9%)		36 (9.1%)	207 (36.6%)	
4	746 (46.4%)	782 (35.0%)		190 (48.2%)	207 (36.6%)	
5	626 (39.0%)	139 (6.3%)		157 (39.9%)	28 (4.9%)	<0.001

Table 2. Patient characteristics. *CsPCa* clinically significant prostate cancer, *PSA* prostate specific antigen, *PV* prostate volume, *PSAD* PSA density, *DRE* digital rectal exam findings, *FH* family history of PCa. *PNB* previous negative biopsy.

elastic net model the best combination of L2 and L1 regularization was reached for 0.4 parameter. The results suggests that PSA can be removed from the linear model.

Classification tree

Figure 2 displays the classification tree. The PI-RADS was the variable that best discriminates, categories 1, 2 and 3 correspond to non-CsPCa. On the second level, for PI-RADS 4 and 5 categories, PSAD > 0.1319 classified patients as CsPCa. On the third level, for PSAD < 0.1319, a PI-RADS 5 category identify patients as CsPCa. In the case of a PI-RADS 4 category, a DRE tagged as non suspicious corresponds to non-CsPCa, by contrast, a DRE tagged as suspicious classified patients as CsPCa or non-CsPCa depending on a PSAD value above or below 0.1014 respectively.

Additionally, to explore the relation between variables and the outcome we used the SHAP values. These values are a tool for understanding the relative significance of features within a predictive model. By assessing the influence of each feature on the model's predictions, SHAP values offer a graphical overview of each variable's impact in PCa prediction. We evaluate SHAP values for the 960 validation data, and subsequently the summary plot show the influence of the variables for each individual in the CsPCa prediction.

Variable	OR (95% C.I.)	p-value
Age (years)	1.059 (1.048,1.071)	<0.001
rcs(PSA, 5) PSA	1.579 (1.355,1.850)	<0.001
rcs(PSA, 5) PSA'	2.150×10^{-7} (1.662×10^{-11} , 2.425×10^{-3})	0.001
rcs(PSA, 5) PSA''	3.449×10^{15} (7.171×10^4 , 2.148×10^{26})	0.004
rcs(PSA, 5) PSA'''	1.111×10^{-10} (7.608×10^{-19} , 1.464×10^{-2})	0.016
PV	0.974 (0.971, 0.977)	<0.001
DRE (suspicious)	2.599 (2.133,3.170)	<0.001
FH (yes)	1.673 (1.227,2.278)	0.001
PNB (yes)	0.696 (0.579,0.835)	<0.001
PI-RADS 2:1	1.279 (0.690,2.301)	0.421
PI-RADS 3:1	1.839 (1.288,2.662)	<0.001
PI-RADS 4:1	6.755 (4.882,9.520)	<0.001
PI-RADS 5:1	18.758 (12.994,27.515)	<0.001

Table 3. Multivariate logistic regression model. *rcs* restricted cubic spline, *PSA* prostate specific antigen, *PV* prostate volume, *PSAD* PSA density, *DRE* digital rectal exam findings, *FH* family history of PCa, *PNB* previous negative biopsy.

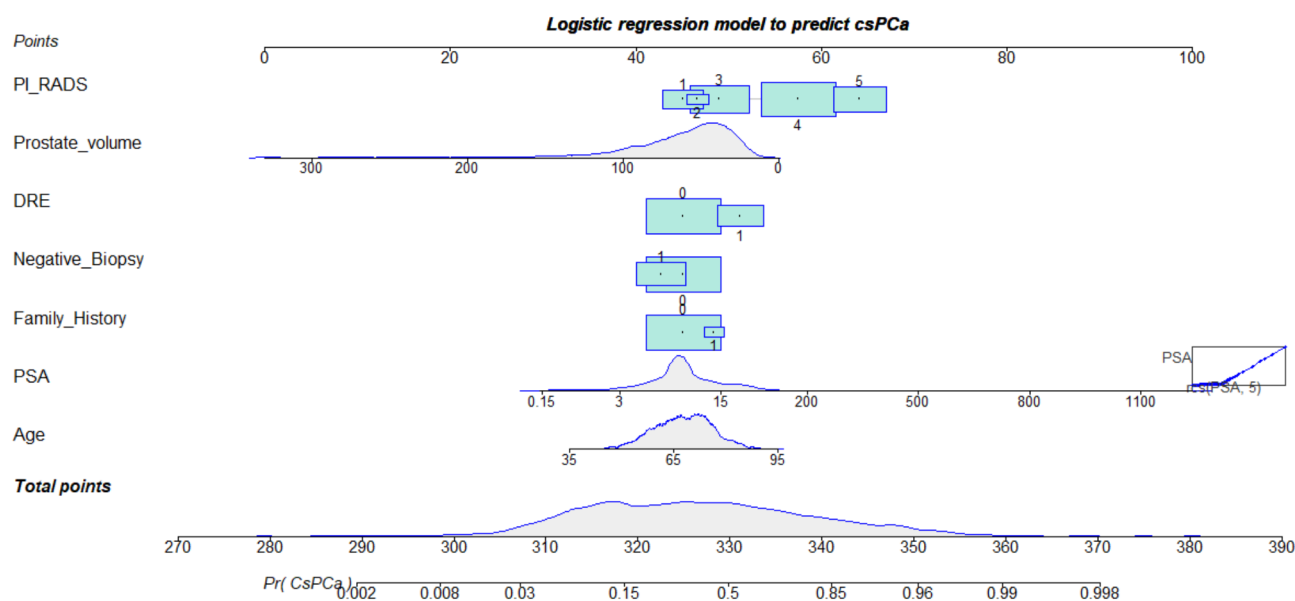


Fig. 1. Nomogram for the multivariate logistic regression model.

In Fig. 3 we shown that the three variables that exerted the most influence on CsPCa prediction were the PI-RADS, PSAD and DRE.

Random forest

We added a total of 300 trees to the ensemble, as this number was found to be reasonable based on the reduction in prediction error depicted in (Fig. 4). The results of the hyperparameter optimization are depicted in (Fig. 5). The best-performing model was achieved with a minimum size of 8 for the terminal node and training each tree with 7 variables.

Figure 6 illustrates the variable importance in the random forest, highlighting the most significant variables. It is noteworthy that in a more robust model like the random forest, the importance of the PI-RADS probability becomes increasingly prominent to predict CsPCa.

Extreme Gradient boosting

Regarding parameters used, best model was attained using 800 trees, a learning rate of 0.2, a max depth of four levels in the trees, and a subsample of 80% of training data and 80% of predictor variables. Also, early stopping was set to 80 and we used L2 regularization.

Figure 7 shows the estimated SHAP summary values for the validation data. The chart shows the importance of each variable for each individual as a point on the graph. Darker values represent higher influence. It can

Variable	Ridge regression	LASSO	Elastic net
Intercept	-6.175741229	-7.39505896	-5.923857
Age	0.054093253	0.06274349	0.004972360
PSA	0.005507262	0.00000000	0.000006632693
PSAD	0.971476782	2.16508376	0.9623193
PV	-0.015923642	-0.01723216	-0.01447082
DRE	0.779058970	0.83286862	0.7345436
FH	0.350414930	0.44284927	0.1688009
PNB	-0.246643897	-0.27203604	-0.1583635
PIRADS	0.737006746	0.87783219	0.7411823

Table 4. Ridge regression, LASSO and elastic net coefficients. *PSA* prostate specific antigen, *PV* prostate volume, *PSAD* PSA density, *DRE* digital rectal exam findings, *FH* family history of PCa, *PNB* previous negative biopsy.

be observed that for a group of individuals, the variable PI-RADS is the most influential as a risk factor, corresponding to individuals with PI-RADS 4 and 5. It can also be noted that for other individuals, having a positive DRE or family history of PCa are highly influential. Conversely, having a previous negative biopsy is also very influential in other individuals, but as a protective factor.

Neural networks

The best model was attained using an architecture as 8-8-1, indicating 8 input nodes, 8 nodes in the hidden layer, and 1 output node. A total of 81 weights were estimated, and the hyperbolic tangent activation function was used. The optimization function employed was cross-entropy, which measures the disparity between predictions and the actual occurrence of CsPCa. The architecture of the network is visually represented in Fig. 8, with positive weights depicted as black lines and negative weights shown as grey lines. The thickness of each line corresponds to the relative magnitude of the weight it represents.

Figure 9 illustrates the variable importance plot for the multilayer perceptron, utilizing the SHAP values. The variables that exerted the greatest influence were very similar to the XGBoost model, PI-RADS, DRE, a previous negative biopsy and a family history of PCa.

Validation of developed models

To evaluate the validity of the models, we utilized a separate set of validation data comprising 960 patients, a 20% of the total dataset. This approach allowed us to assess the performance of the machine learning models on data that were not used during the model development phase.

Regarding the probabilities provided by the models, we presented their distribution in a comparative boxplot in (Fig. 10). All models exhibited good discrimination ability, with the outcomes of Ridge regression, LASSO, and Elastic Net normalized for comparability.

Regarding the discrimination ability, in Table 5 can be seen that all models exhibit good discrimination with AUC values above 0.8. The highest AUC is achieved by the XGBoost and the neural network that show equivalence between both ($p=0.981$) with an AUC of 0.872 (0.849, 0.895), which shows no statistically significant difference compared to logistic regression (0.870 (0.848, 0.893), $p=0.925$ and $p=0.930$) and the random forest models (0.860 (0.836–0.884), $p=0.485$ and $p=0.490$). However, both models demonstrates superiority over the Ridge regression (0.854 (0.831, 0.879), $p=0.007$ and $p=0.002$), LASSO (0.857 (0.834, 0.881), $p=0.017$ and $p=0.007$), Elastic net (0.857 (0.834,0.881), $p=0.018$ and $p=0.008$) and classification tree models (0.807 (0.780,0.835), $p<0.001$ and $p<0.001$). The ROC curves are displayed in (Fig. 11).

Although we observed similar behavior in terms of AUC values, it is crucial for the predictive models to be effective in detecting CsPCa cases, especially at high sensitivity values. Table 6 summarizes the specificities for high sensitivity values. When considering a sensitivity value of 0.9, the XGBoost model performs the best with a specificity of 0.640. It is followed by the random forest (0.638, $p=0.999$), neural network (0.634, $p=0.901$) and logistic regression (0.620, $p=0.538$).

In order to effectively prioritize the identification of CsPCa cases, our study emphasizes logistic regression, random forest, XGBoost and one hidden layer perceptron as the most optimal models. However, a crucial question remains: to what extent can these models help reduce the number of unnecessary biopsies? This important aspect can be examined by analyzing the clinical utility curves presented in (Fig. 12a,b). The X-axis represents the potential threshold CsPCa probability points used to classify individuals as CsPCa or non-CsPCa. On the Y-axis, we present the percentage of CsPCa cases misclassified below the selected cut-off point (indicated by a solid line) and the percentage of biopsies that could be avoided (depicted with a dotted line).

By analyzing the clinical utility curve, we can determine the number of biopsies that could be avoided by detecting a certain percentage of CsPCa cases. Table 7 presents the percentage of biopsies avoided for a wrongly CsPCa classification rate. For example, when there is a 10% misclassification rate of CsPCa cases, the XGBoost is able to avoid 41.77% of unnecessary biopsies. The random forest and neural networks closely follows with rates of 41.67% ($p=0.999$) and 41.46% ($p=0.931$) respectively, while the logistic regression model have a lower rate of 40.62% ($p=0.641$), but with no statistical significant differences between models.

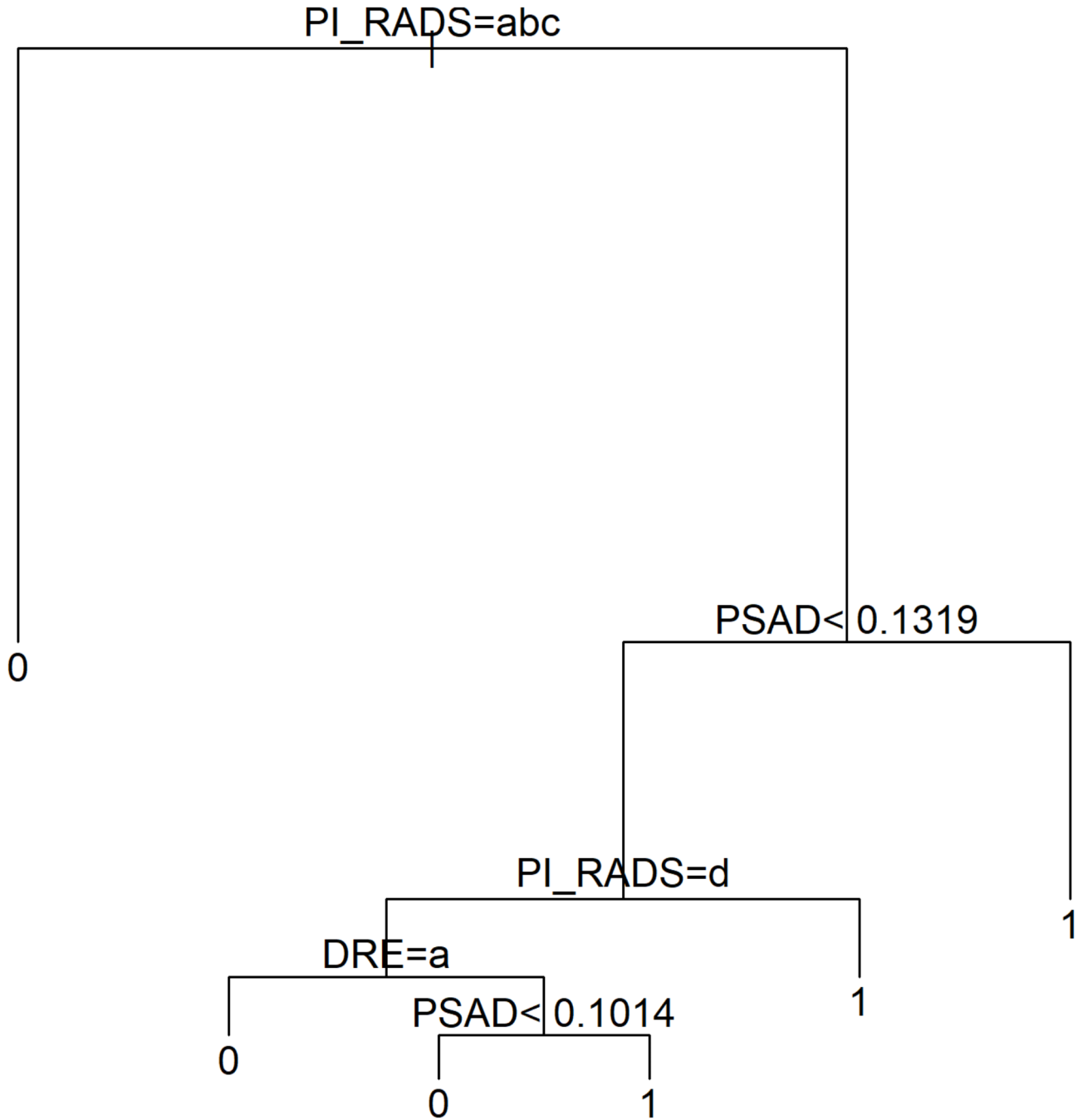


Fig. 2. Classification tree. Class 1: CsPCa, class 0: non-CsPCa. PI-RADS a:1,b:2, c:3, d:4, e:5, DRE a: suspicious.

Discussion

Machine learning techniques have exhibited superior predictive capabilities in various medical domains²⁴. Determining the optimal point at which these models become most compelling necessitates a balance between model fit and applicability²⁵. One of the traditional disadvantages attributed to such models is their black-box nature. However, this issue has undergone significant transformation in recent years, with various tools being developed to analyze the importance of variables in prediction²⁶.

Hiremath et al.²⁷ utilized deep learning to combine PI-RADS scores and clinical variables for csPCa detection on biparametric MRI, achieving an area under the curve (AUC) of 0.810. Similarly, Wang et al.²⁸ introduced a dual-path convolutional neural network (CNN) that reported sensitivities of 0.6374 and 0.8978 at 0.1 and 1 false positives per patient, respectively. These studies demonstrate the potential of deep learning to optimize the diagnostic performance of MRI by integrating imaging features with contextual clinical data.

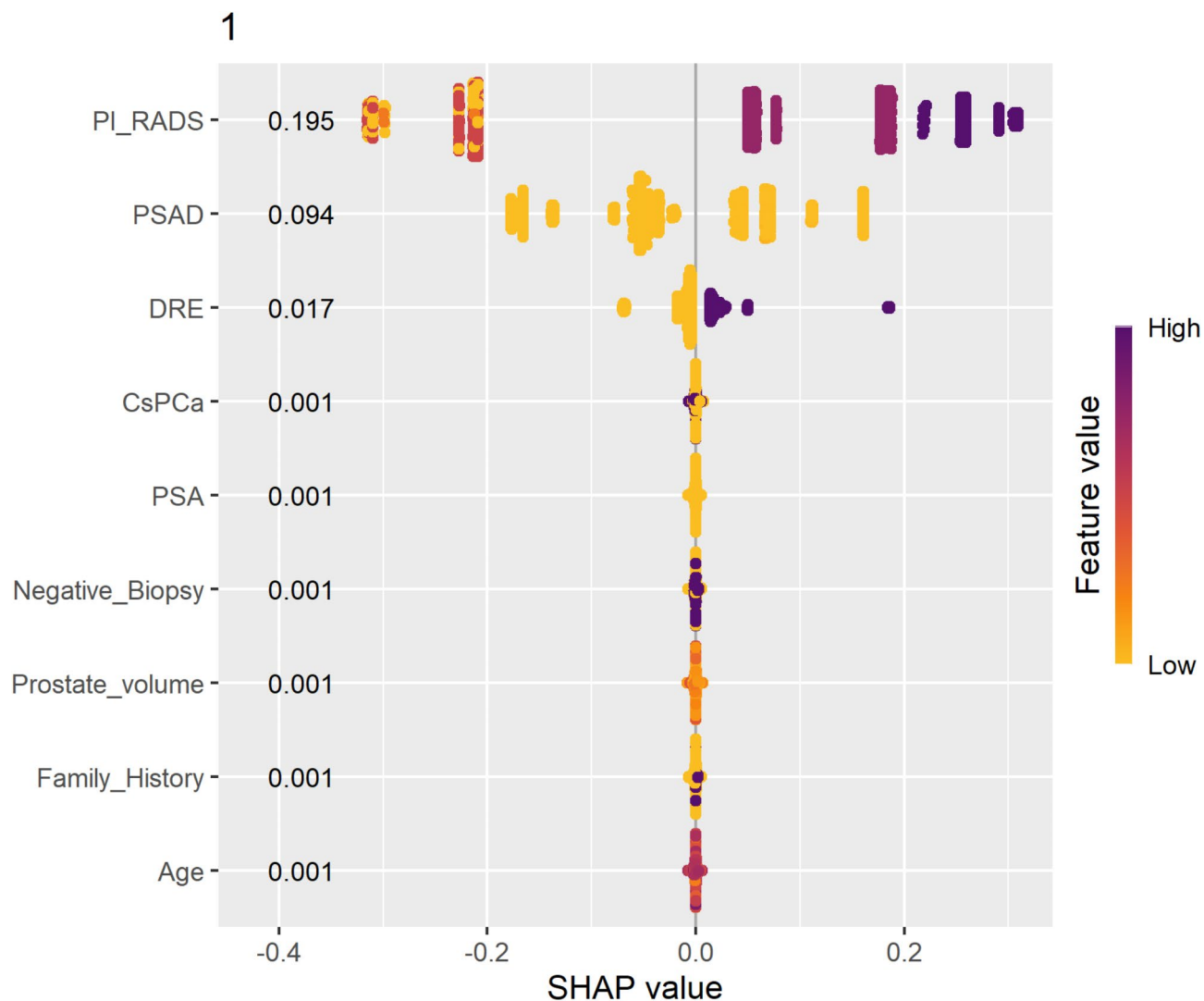


Fig. 3. Variable importance through SHAP analysis in classification tree.

Combining MRI features with clinical parameters can significantly enhance the predictive accuracy of AI models. Peng et al.²⁹ trained machine learning models, including logistic regression, decision trees, and random forests, that incorporated multiparametric MRI-derived texture analyses, dynamic contrast-enhanced MRI, and clinical parameters such as PSA density. The random forest model achieved the highest performance with an AUC of 0.925. Despite promising results, their study was limited by a small sample size (194 patients), underscoring the need for larger datasets.

Donisi et al.³⁰ extended this approach by combining radiomic features and clinical variables, employing decision trees, random forests, gradient boosting trees, AdaBoost, Naïve Bayes, and k-nearest neighbors. The gradient boosting tree model achieved the best AUC of 0.774 for distinguishing csPCa from clinically insignificant prostate lesions. While these approaches highlight the added value of integrating clinical data, they lack comprehensive evaluation of their clinical utility and real-world applicability.

A predictive model integrating PSAD, PI-RADS score, and age achieved an AUC of 0.938, outperforming individual components (e.g., PI-RADS alone had an AUC of 0.875). This model showed significant clinical utility in decision curve analysis, especially for reducing unnecessary biopsies. External validation confirmed its robustness with high sensitivity and specificity^{31,32}.

Studies compared semi-quantitative (e.g., PI-RADS scores) and quantitative MRI parameters (e.g., ADC, T2WI, Ktrans, Kep) with PSAD. PI-RADS combined with PSAD consistently improved diagnostic performance for csPCa. Logistic regression models incorporating these parameters demonstrated high predictive power, with ROC curve analysis ranking PSAD among the top contributors^{32,33}.

Machine learning methods integrating PSAD and MRI features, such as random forests and support vector machines, have shown promise in improving diagnostic accuracy. For example, models including PSAD and mpMRI features achieved AUC values up to 0.925, validating their utility in external datasets^{31,32}. Combining PSAD with MRI metrics refines risk stratification, reducing unnecessary biopsies and focusing interventions on

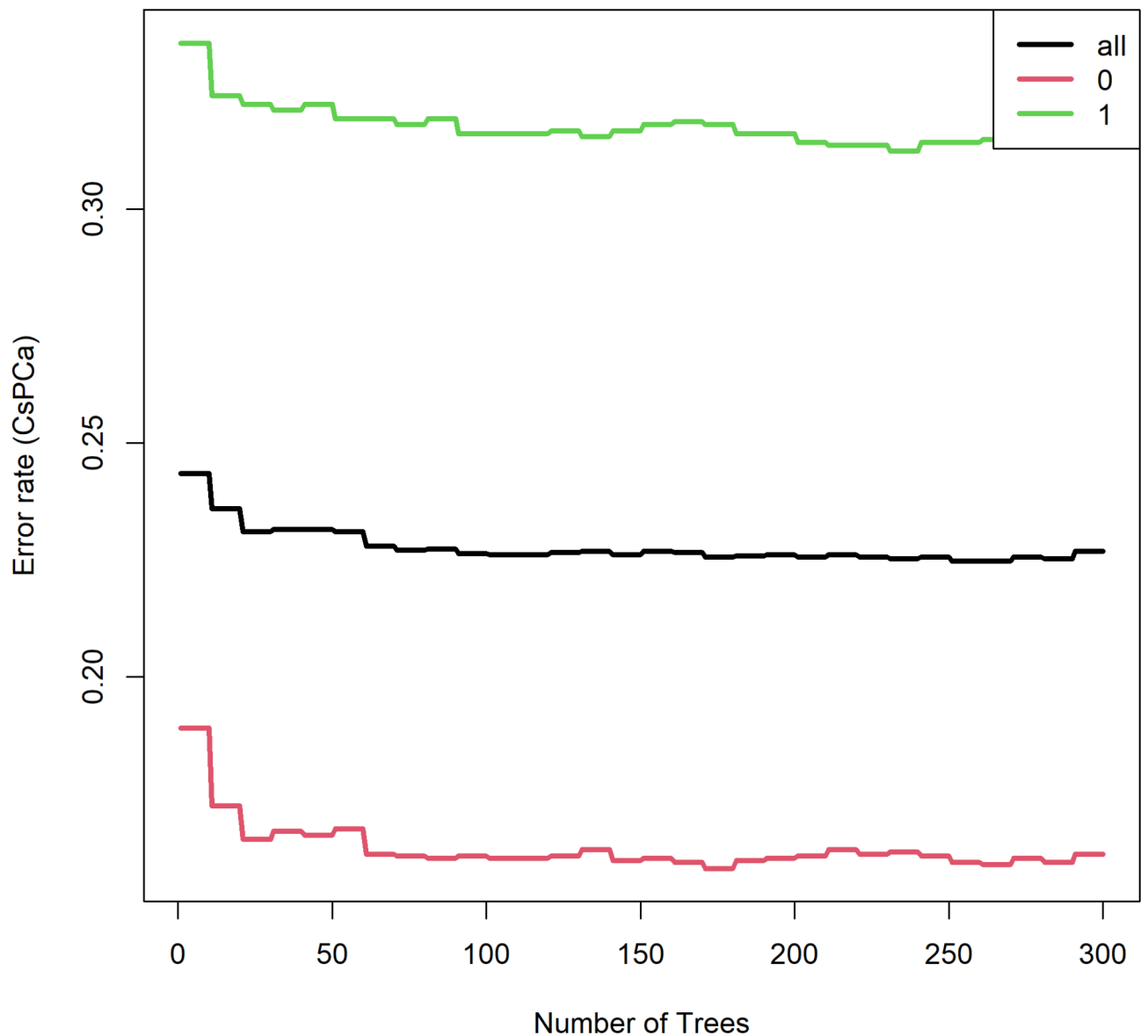


Fig. 4. Minimization of error depending on the number of trees added to the model. The error rate was evaluated for the non-CsPCa cases (0), CsPCa cases (1), and all data, using the out-of-bag (OOB) error estimation.

patients with higher probabilities of csPCa. A significant challenge lies in standardizing MRI acquisition protocols, PSAD thresholds, and ensuring external validation to generalize findings across diverse populations^{32,33}.

Previously, we had developed a logistic regression model for predicting CsPCa using mpMRI PIR-RADS report¹⁴, which has been implemented into a freely accessible application <https://mripcaprediction.shinyapps.io/MRIPCaPrediction/>. Furthermore, it has been externally validated, demonstrating good performance in the Catalonia region. The developed model could serve as a valuable tool for decision-making regarding treatment, particularly concerning whether a patient should undergo diagnostic biopsy to determine the Gleason score³⁴.

In this study, we aimed to analyze predictive enhancement through the utilization of machine learning algorithms. To achieve this, we employed regularization techniques, tree ensemble methods, and neural networks. Regarding discrimination capacity assessed by the AUC, findings demonstrated remarkable similarity among logistic regression (0.870), XGBoost (0.872), and neural network (0.872) models. Although the AUC is the most commonly used parameter for validating a predictive model, in a binary classification problem, it is essential to choose a cutoff point to classify individuals. In this regard, a cutoff point is represented by a single pair (Sensitivity, Specificity) on the ROC curve. Therefore, the most critical aspect for the clinical use of a model lies in selecting a criterion to determine this cutoff point. Misclassifying a case of CsPCa is a significant loss because it corresponds to prostate cancer with an unfavorable prognosis. The criterion must thus be stringent; in this case, it could be set with a false positive rate of 10%, or equivalently, with a sensitivity of 90%. Consequently,

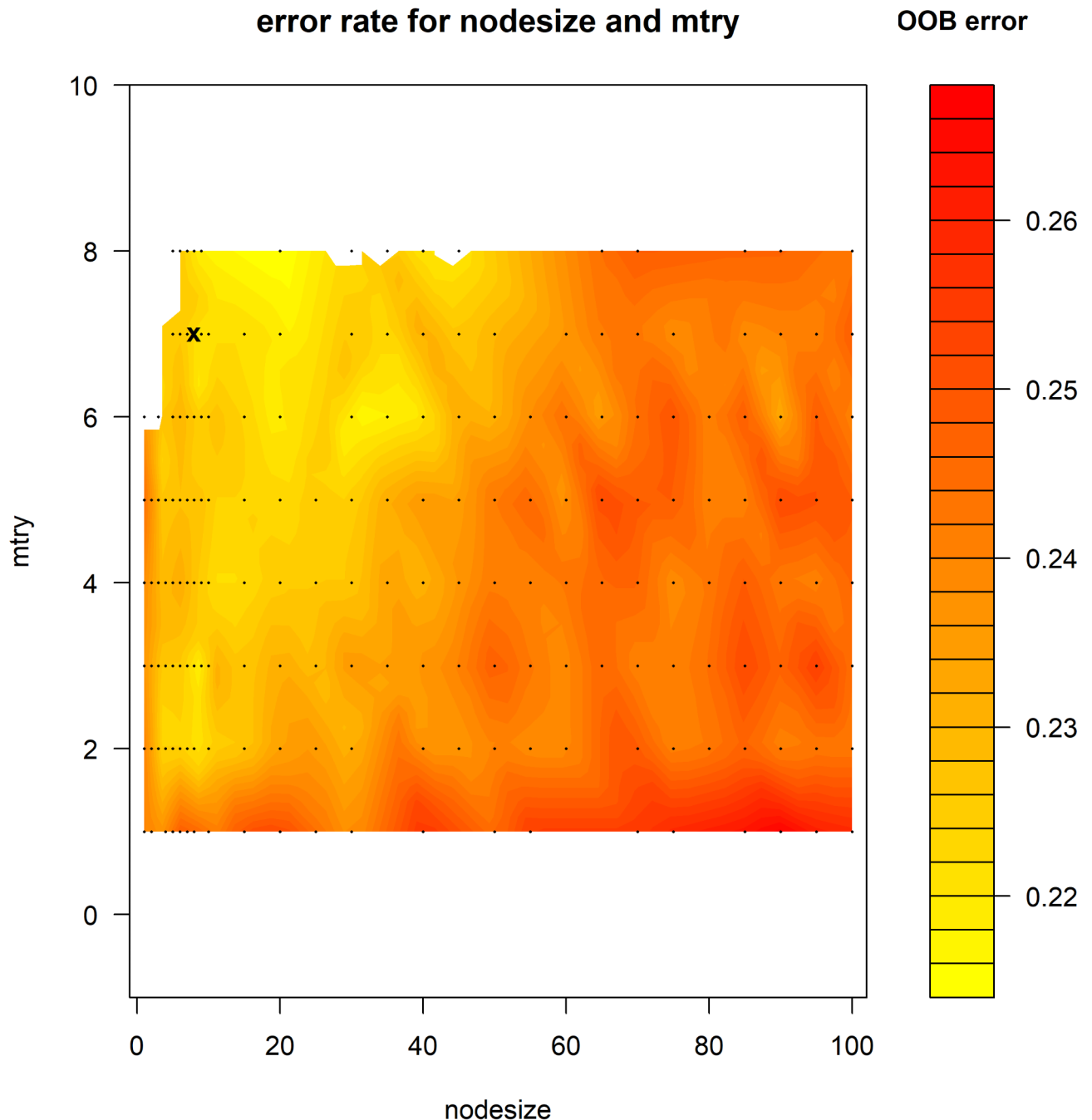


Fig. 5. Minimization of error depending on the nodesize (minimum size of terminal node) and mtry (number of variables to possibly split at each node).

our interest lies in the model that achieves the highest specificity for a sensitivity of 90%. The models exhibiting the best performance in this scenario are random forest (0.638), XGBoost (0.640), and neural networks (0.634).

Additionally, it is essential to ascertain the clinical utility of the models. In a screening scenario combining clinical variables and mpMRI reporting, it is crucial to determine the proportion of patients who can be classified as low risk and thus avoid undergoing biopsies. Although prostate cancer (PCa) screening reduces the incidence of advanced disease and mortality, trade-offs include overdiagnosis and resultant overtreatment³⁵. Nordström et al. have reported diagnostic outcomes from repeated prostate cancer screening utilizing magnetic resonance imaging for men with elevated serum PSA levels, lesion-targeted prostate biopsies, and risk-stratified screening intervals³⁶. Key findings indicate a notable proportion of diagnosed men had potentially CsPCa, low detection rates of low-grade (grade group 1) cancer, and a minimal occurrence of suspicious lesions for cancer on repeated MRI scans. Hence, as Bratt suggests, the aim is to create a risk-stratified model for prostate cancer screening that can effectively decrease prostate cancer mortality while maintaining acceptable levels of costs and minimizing harms associated with diagnostic resource allocation, overdiagnosis, and overtreatment³⁷.

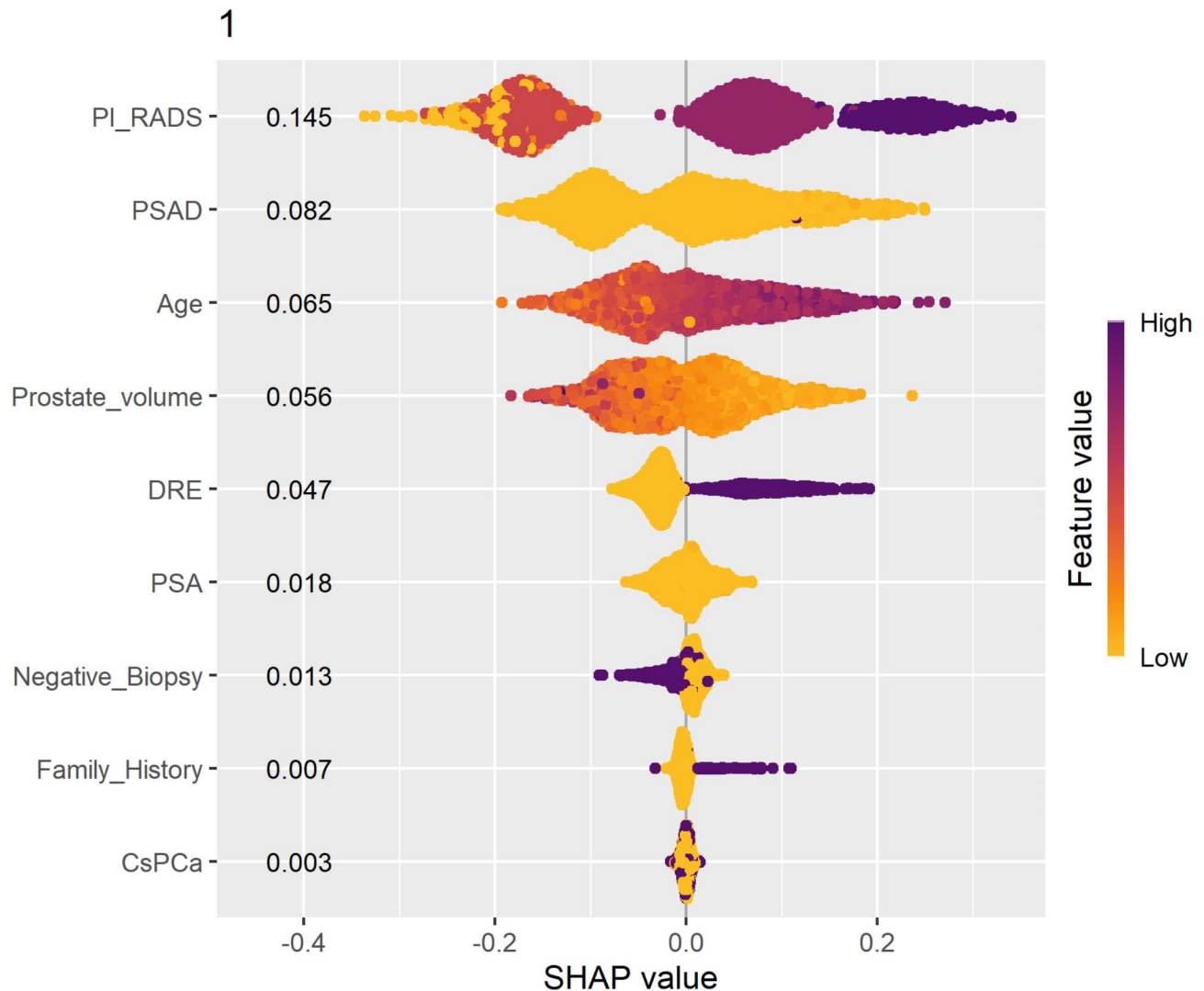


Fig. 6. Variable importance through SHAP analysis in random forest model.

In this aspect, we believe our work is particularly innovative. While previous machine learning models for predicting clinically significant prostate cancer (CsPCa) exist, they often report results in terms of discrimination or accuracy alone. Demonstrating clinical utility is essential for the routine clinical application of a prostate cancer model. High prediction loss percentages cannot be accepted, as clinically significant cancer may be the one with the worst prognosis. In this regard, a loss greater than 10%, or a sensitivity below 90%, seems unacceptable. These high sensitivity values may correspond to low specificity or involve very few patients below the cutoff, meaning the model lacks clinical utility, as it would lead to biopsies for almost the entire population. Estimating the percentage of biopsies avoided with a lower classified CsPCa rate is crucial. Our analysis shows that for a diagnostic loss of 10% in CsPCa, machine learning models exhibit significant clinical utility, reducing biopsy rates by 41.7, 41.8, and 41.7% for random forest, XGBoost, and neural network models, respectively. Additionally, the logistic regression model also reduces biopsy rates notably by 40.6%. When compared to a simpler protocol, where all patients with PI-RADS greater than or equal to 3 undergo biopsy, the validation database would result in 14% of patients undergoing biopsy. Therefore, a reduction in biopsy rates of approximately 26–27% is observed, depending on the model applied.

Finally, we analyzed the importance of variables in the models both globally and through individualized analysis using Shapley values. The widespread use of logistic regression models for binary prediction in prostate cancer is evident due to their ability to generate odds ratios and visualize variable roles in nomograms³⁸. Machine learning models have also developed methods to visualize variable importance in prediction, although the role of each variable in prediction was not visualized until the proposal of using Shapley values.

The computation of Shapley values originates from game theory and serves as a method for allocating payments in cooperative games by analyzing the coalitions that can be formed among players and the contribution of each player when added to the coalition³⁹. This concept has been applied to analyze the contribution of a variable to a prediction model by considering its addition to another set of predictor variables. The estimation can be

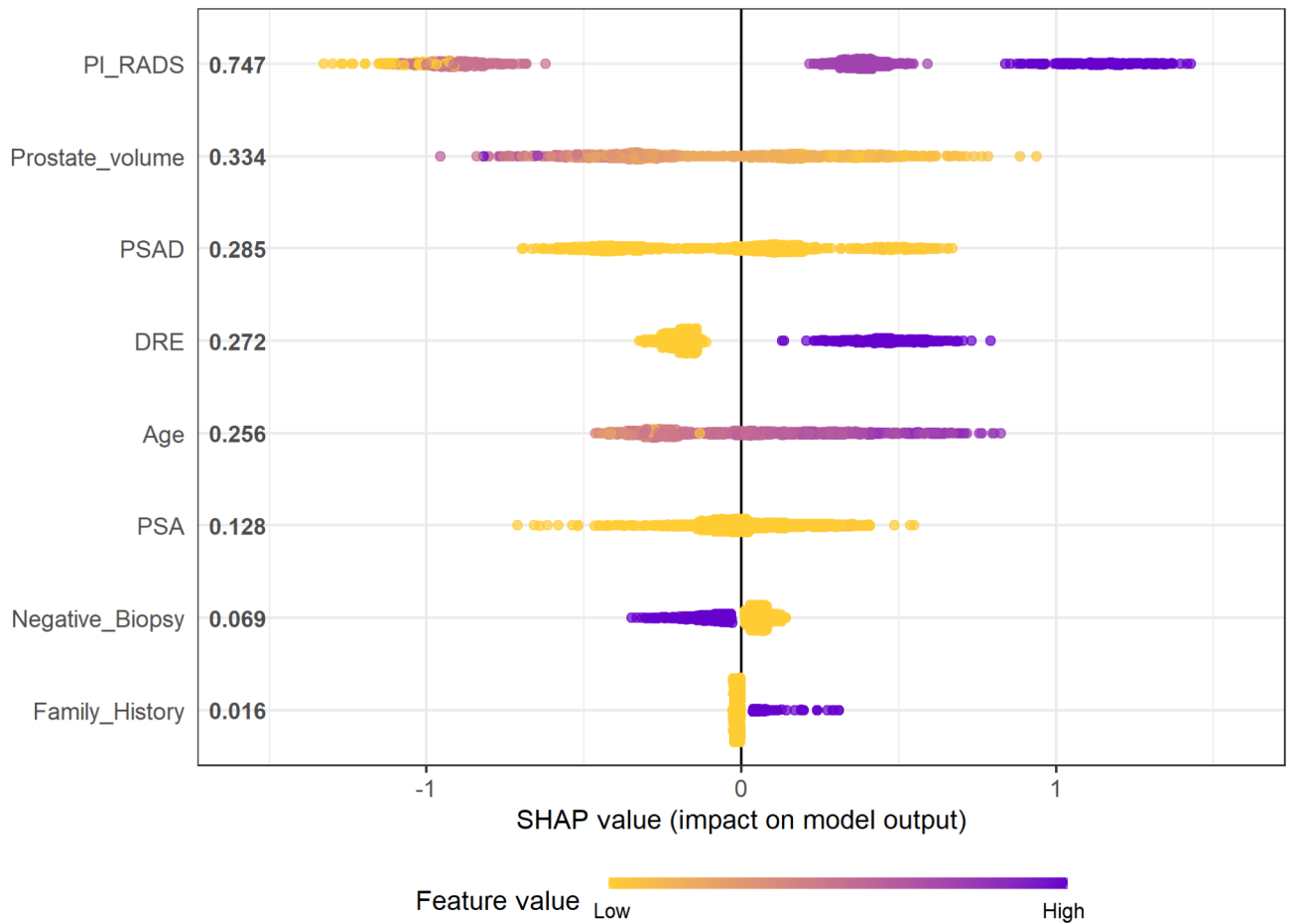


Fig. 7. SHAP summary plot for XGBoost model.

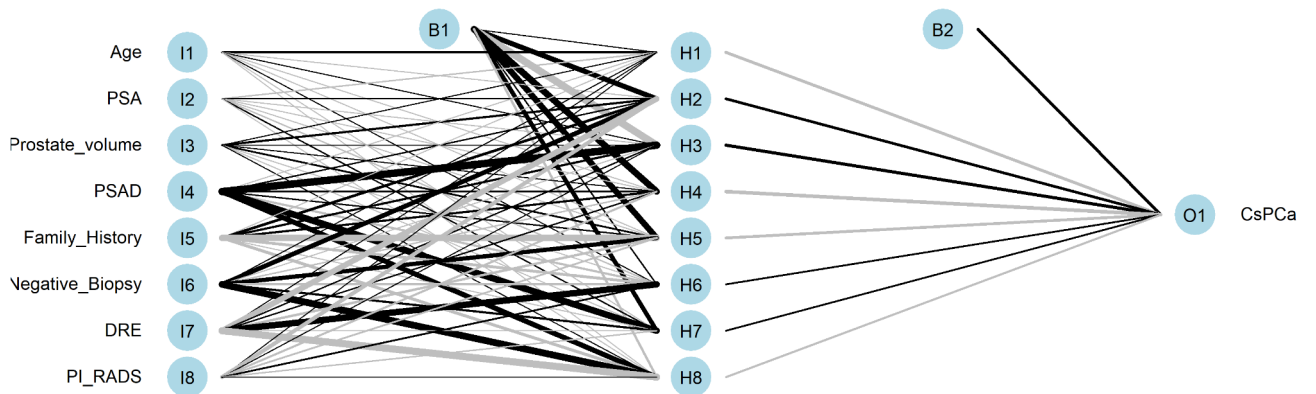


Fig. 8. Neural network architecture with input (I), hidden (H), and output (O) layers. (B) is the result obtained after applying the activation function.

performed for each individual and subsequently visualized collectively for all cases. To illustrate its performance we shown in (Fig. 13) three different cases for the neural network model.

The first case pertains to a low-risk patient that was non CsPCa, with the neural network predicting a probability of 0.102 for prostate cancer. This patient is a 71-year-old individual with a PI-RADS score of 2, a PSA level of 3.9 ng/mL, a prostate volume of 40 mL, undergoing initial biopsy, devoid of familial history, and exhibiting no suspicious findings upon DRE. For this patient, the model elucidates that the most influential factor in predicting non-CsPCa is the PI-RADS score 2 in the report, which significantly outweighs other factors (−0.255) as a protective factor. Additionally, low values of PSAD, DRE, or serum PSA level contribute to this

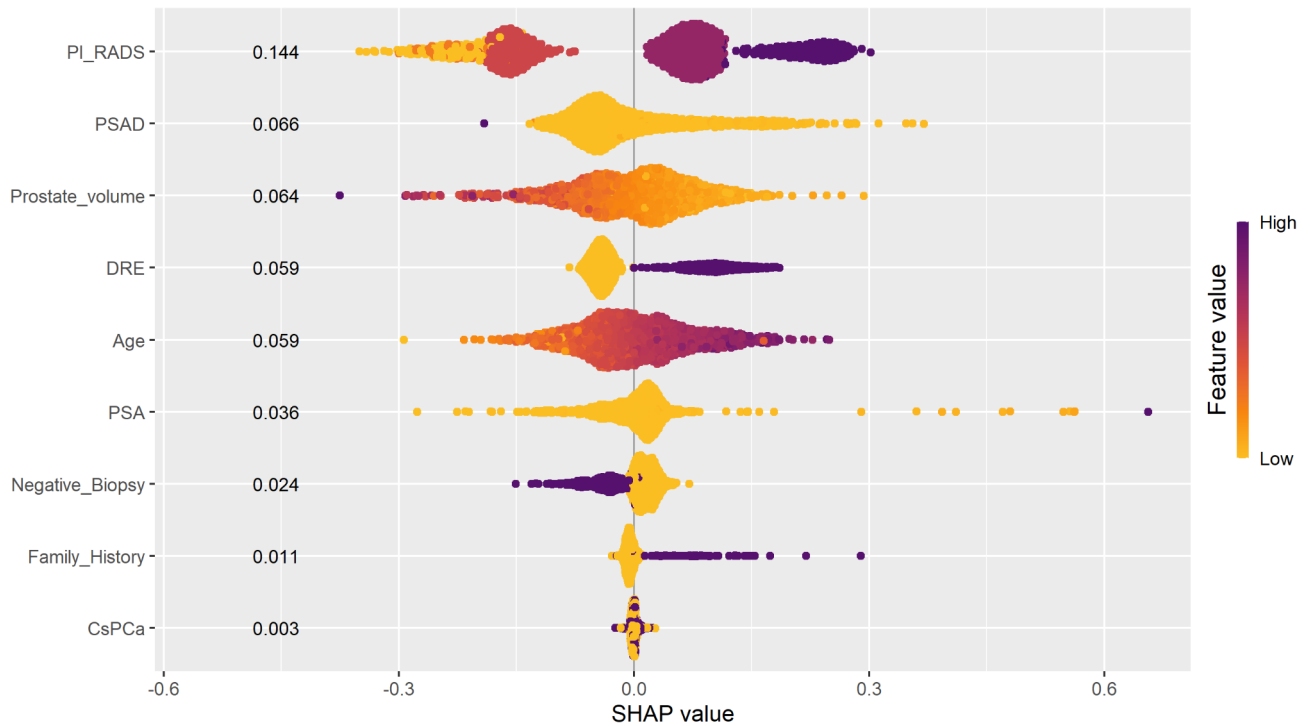


Fig. 9. SHAP values in the neural network.

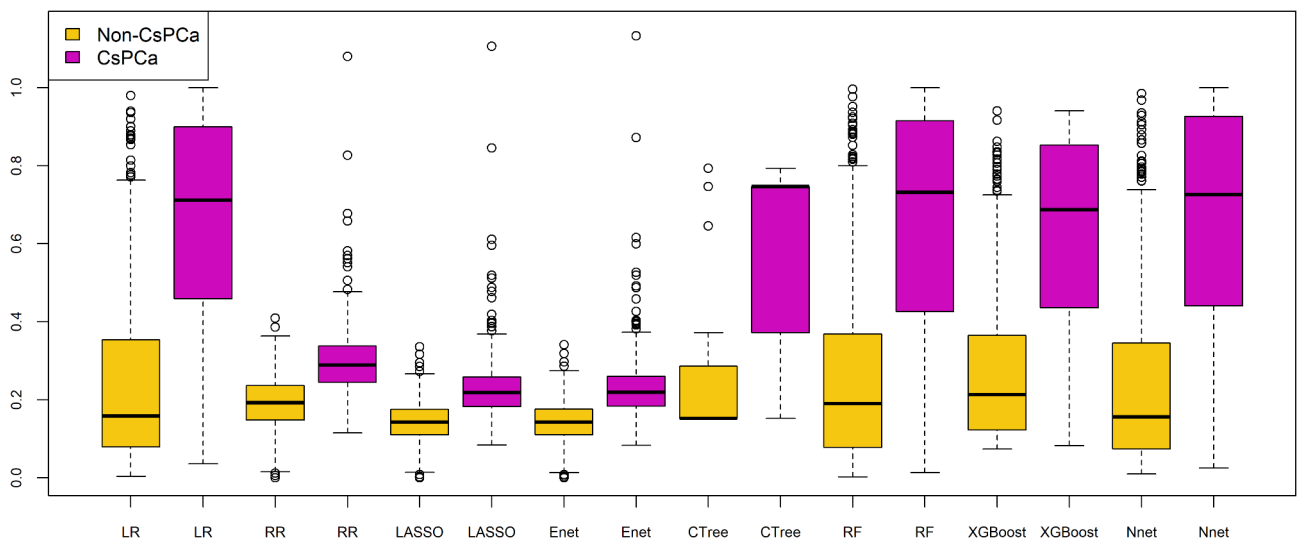


Fig. 10. Probabilities distributions. *LR* logistic regression, *RR* ridge regression, *Enet* elastic net, *CTree* classification tree, *RF* random forest, *XGBoost* extreme gradient boosting, *Nnet* neural network.

prediction, albeit to a lesser extent, while factors considered as risk indicators demonstrate minimal importance. The second case involves a 67-year-old patient with a PI-RADS score of 3, a PSA level of 21 ng/mL, and a prostate volume of 207 mL, having had a previous negative biopsy, no familial history, and exhibiting non-suspicious findings on DRE, ultimately resulting in a non-CsPCa diagnosis. Analysis using SHAP reveals that the most critical factor for prediction is prostate volume, followed by an intermediate PI-RADS score of 3, while the remaining variables demonstrate limited significance. The third case involves a high-risk patient, a 56-year-old individual with a PSA level of 7.3 ng/mL, a prostate volume of 20 mL, no familial history or previous negative biopsies, exhibiting a suspicious lesion on DRE and a PI-RADS score of 5, resulting in CsPCa. SHAP analysis in this case indicates that the PI-RADS score of 5 is the most decisive factor for high-risk categorization, followed by PSAD, with DRE results having a lesser impact.

	LR	RR	LASSO	Elastic net	CT	RF	XGBoost	NN
	0.870 (0.848–0.893)	0.854 (0.831– 0.879)	0.857 (0.834– 0.881)	0.857 (0.834–0.881)	0.807 (0.780– 0.835)	0.860 (0.836– 0.884)	0.872 (0.849–0.894)	0.872 (0.849– 0.894)
LR		0.359	0.438	0.444	<0.001	0.042	0.925	0.930
RR			0.011	0.008	0.010	0.760	0.007	0.002
LASSO				0.185	0.007	0.869	0.017	0.007
Elastic net					0.006	0.876	0.018	0.008
CT						<0.001	<0.001	<0.001
RF							0.485	0.490
XGBoost								0.981

Table 5. Area under the ROC curve for the developed models on validation data. *LR* logistic regression, *RR* ridge regression, *CT* classification tree, *RF* random forest, *XGBoost* extreme gradient boosting.

These values illustrate how a neural network can assess the impact of variables on prediction for each individual. They also demonstrate how all cases can be collectively evaluated, as depicted in Fig. 9, where all individuals are presented together. This allows for a comprehensive analysis of the variables' collective influence on prediction.

There are still challenges to overcome in order to apply MRI models in prostate cancer prediction, one of them is the variability among radiologist. Studies have shown that radiologists can have varying levels of agreement when assigning PI-RADS scores, particularly for intermediate scores (PI-RADS 3), which are more difficult to interpret⁴⁰. This variability arises due to differences in experience, training, and interpretation of subtle MRI findings. It has been found that intra- and inter-observer variability is especially evident when interpreting prostate lesions in complex anatomical regions or when distinguishing between benign and malignant findings in low-grade lesions. Inconsistent PI-RADS scoring can lead to differences in patient management, particularly in decisions regarding biopsy. A high PI-RADS score might prompt a biopsy in one instance, while a lower score could lead to watchful waiting in another. These discrepancies can result in missed diagnoses or unnecessary biopsies, both of which have significant clinical and financial implications⁴¹.

AI models, particularly deep learning algorithms, have shown promise in addressing the subjectivity and variability associated with PI-RADS scoring⁴². By training on large datasets of annotated prostate MRI scans, AI models can assist in providing more consistent and reproducible scores⁴³. These models can standardize the interpretation of prostate MRI findings, reducing the variability between radiologists and increasing diagnostic accuracy⁴⁰. AI-driven approaches can improve diagnostic performance by focusing on the most clinically relevant features of MRI images that might be missed by human interpreters. For instance, convolutional neural networks (CNNs) have demonstrated the ability to learn complex patterns in imaging data, which may be difficult for radiologists to interpret without the assistance of computational tools. AI models can also flag areas of uncertainty, prompting radiologists to reconsider specific aspects of the image. Rather than replacing radiologists, AI tools are intended to serve as decision support systems, enhancing human expertise. In this context, AI can act as a second opinion, providing a consistent reference point and reducing the cognitive load on radiologists, especially in high-pressure or time-constrained environments.

Unfortunately, in many low- and middle-income countries (LMICs), there is a significant shortage of radiologists and trained medical professionals who can interpret complex imaging like prostate MRIs. AI-based tools can help fill this gap by offering support to radiologists, especially in remote or underserved areas⁴⁴. The high cost of medical equipment, including MRI machines, and the lack of financial resources to train and employ skilled radiologists in LMICs can hinder effective prostate cancer diagnosis⁴⁵. Implementing AI tools in LMICs involves overcoming several barriers, including limited access to the necessary infrastructure, such as high-performance computing and reliable internet connections⁴⁶. Additionally, healthcare systems in these countries may lack the digital infrastructure to support AI integration into clinical workflows. Another challenge is the need for local adaptation of AI models to ensure that they are culturally relevant and trained on datasets representative of the local population. AI tools could play a crucial role in enhancing the skills of local healthcare providers. For example, AI could be integrated into training programs to help radiologists and clinicians better understand complex diagnostic criteria like PI-RADS. This would help raise the overall quality of prostate cancer care in LMICs. AI solutions have the potential for rapid scaling across a wide range of healthcare settings in LMICs. Once trained, AI models can be deployed on a large scale without the need for extensive infrastructure or human resources, which is a significant advantage in regions with limited healthcare personnel⁴⁷. AI tools can also contribute to large-scale screening programs aimed at early prostate cancer detection, improving outcomes in populations that might otherwise have limited access to regular screening. With the use of mobile health technologies, AI could be used to screen images remotely, thereby overcoming geographical barriers.

Our study exhibits several limitations. Firstly, the heterogeneity of the data is notable, as we drew information from 10 distinct centers, each characterized by unique features. While random effects could have been employed in mixed linear models to account for these differences, our primary focus remained on the application of machine learning algorithms and their subsequent analysis, rather than incorporating a more refined predictive approach that integrates random effects specific to each center. Secondly, the retrospective nature of our analysis poses a limitation. A prospective analysis typically offers greater uniformity in the measurement of variables. Thirdly, the hyperparameter tuning process was automated to optimize the parameters of machine learning

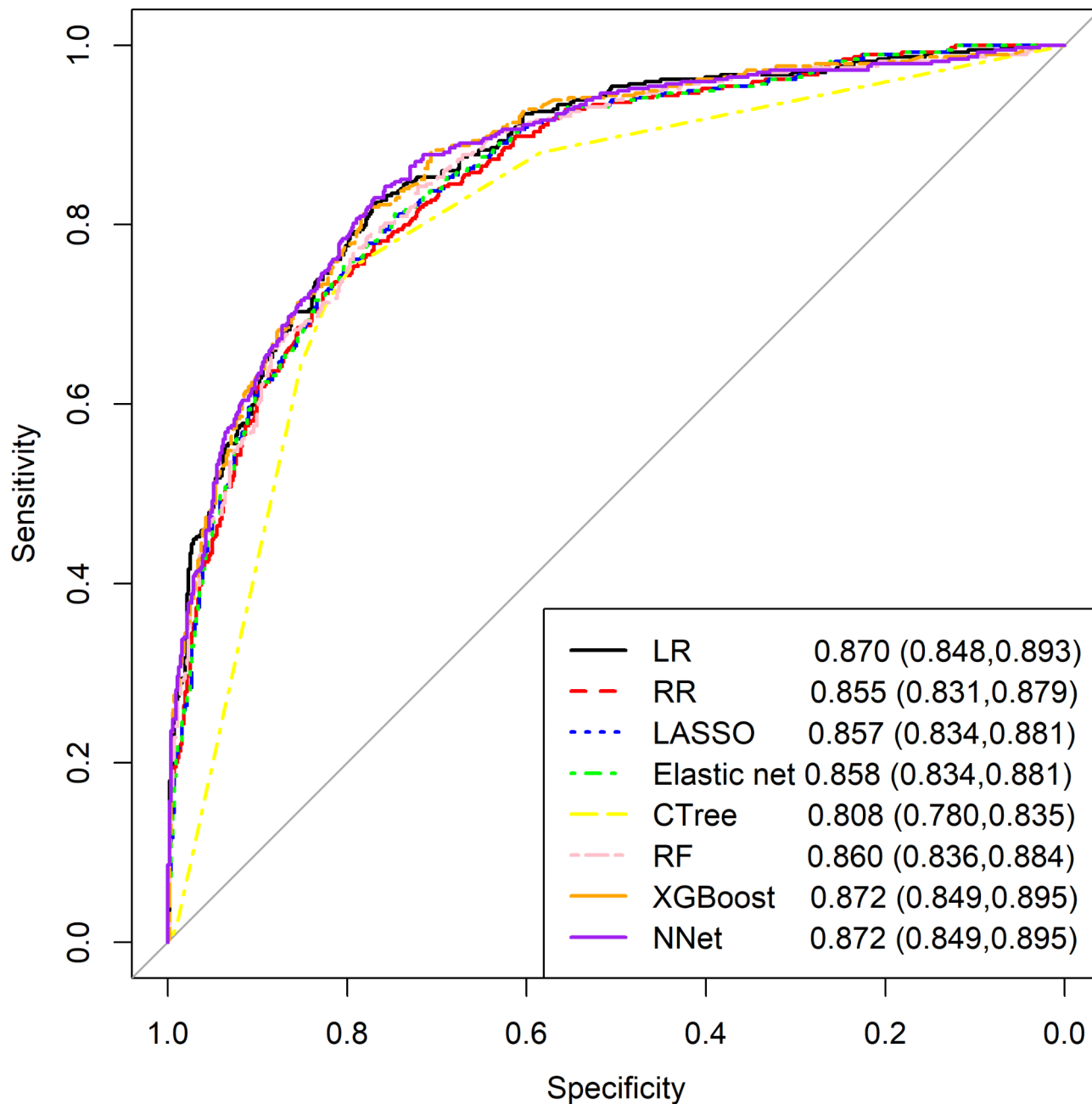


Fig. 11. ROC curves of machine learning models. *LR* logistic regression, *RR* ridge regression, *Enet* elastic net, *CTree* classification tree, *RF* random forest, *XGBoost* extreme gradient boosting, *Nnet* neural network.

models. However, there is potential for improvement through the implementation of a more exhaustive tuning process. Thirdly, to select the cutoff point, we considered a diagnostic CsPCA loss of 5 or 10%, regardless of the patients' PI-RADS categories. This may require refinement, as the prevalence of CsPCA varies significantly across PI-RADS categories⁴⁸.

Conclusions

This study evaluated machine learning models for predicting CsPCA using clinical and MRI data. The XGBoost model, with a 10% false negative rate, reduced biopsy rates by 41.8%. Key factors included PI-RADS categories 4 and 5, suspicious DRE findings, family history, and a negative prior biopsy. The models demonstrated strong predictive power and clear explainability, moving beyond the “black box” approach.

Model	0.80	0.85	0.90	0.95
Logistic regression	0.779 (0.742,0.812)	0.724 (0.685,0.760)	0.620 (0.579,0.660)	0.507 (0.465,0.549)
Ridge regression	0.731 (0.693,0.767)	0.671 (0.631,0.710)	0.592 (0.550,0.632)	0.405 (0.364,0.446)
LASSO	0.751 (0.713,0.785)	0.689 (0.649,0.727)	0.613 (0.571,0.653)	0.392 (0.352,0.434)
Elastic net	0.749 (0.711,0.784)	0.691 (0.651,0.728)	0.611 (0.569,0.651)	0.387 (0.347,0.428)
Classification tree	0.717 (0.678,0.753)	0.635 (0.594,0.675)	0.490 (0.448,0.532)	0.245 (0.211,0.283)
Random forest	0.758 (0.720,0.792)	0.701 (0.661,0.738)	0.638 (0.596,0.677)	0.449 (0.407,0.491)
XGBoost	0.784 (0.748,0.817)	0.717 (0.678,0.754)	0.640 (0.598,0.679)	0.452 (0.411,0.494)
Neural network	0.793 (0.757,0.825)	0.740 (0.702,0.775)	0.634 (0.593,0.674)	0.486 (0.444,0.528)

Table 6. Specificities for sensitivities values provided in each column on validation data. The bold values reflect the best model for the sensitivity value presented in each of the columns.

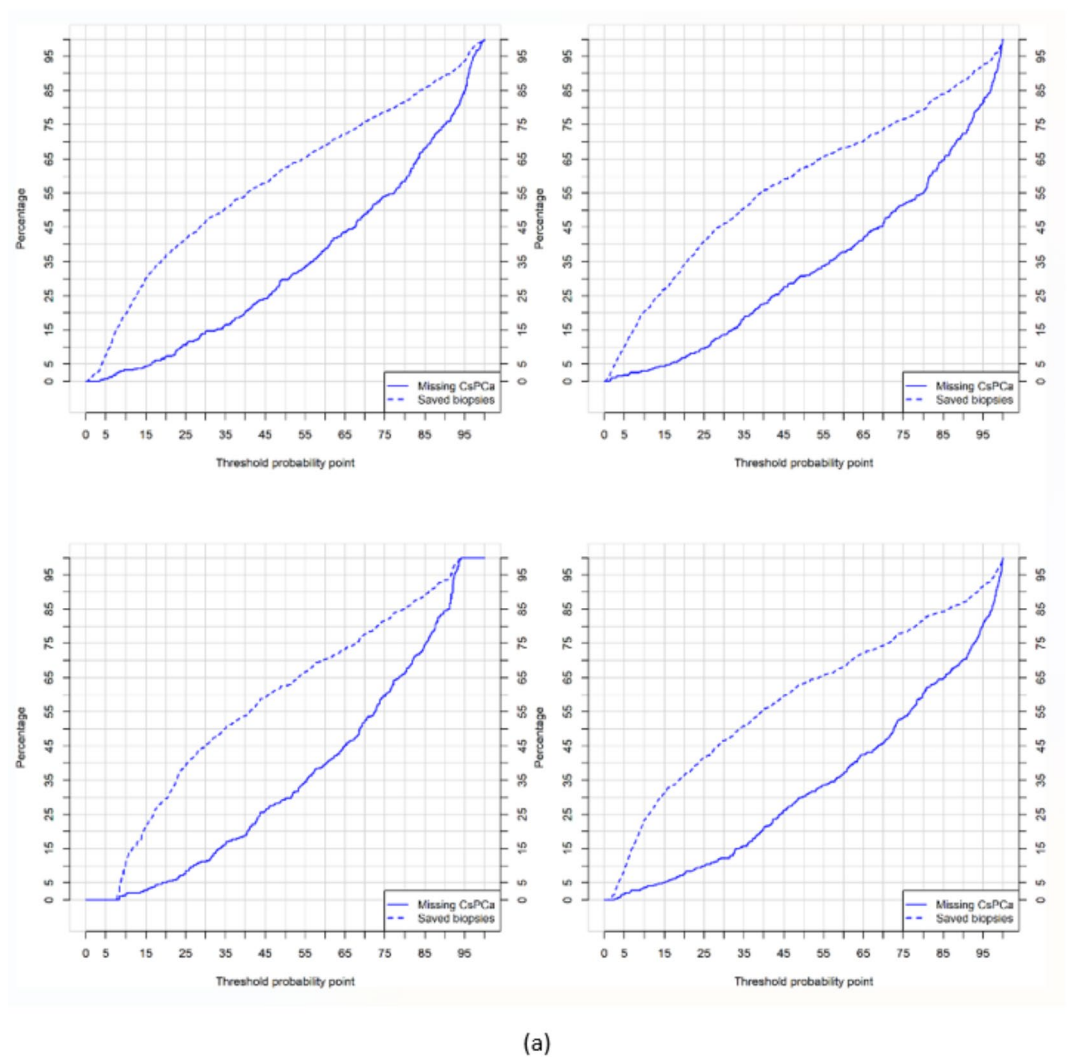
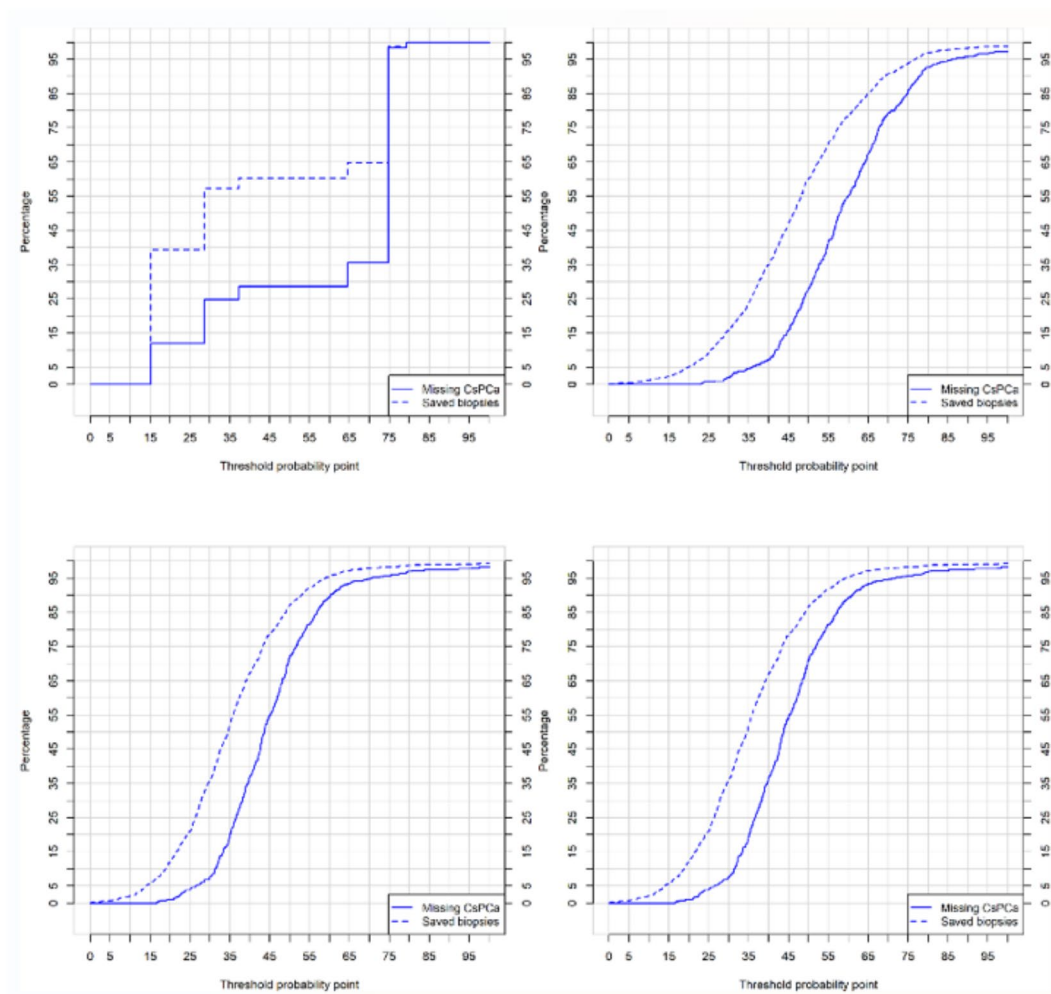


Fig. 12. (a) Clinical utility curves. Upper panel: logistic regression (left), random forest (right), lower panel: XGBoost (left), neural network (right). (b) Clinical utility curves. Upper panel: classification tree (left), ridge regression (right), lower panel: LASSO (left), elastic net (right).



(b)

Figure 12. (continued)

Missing CsPcCa (%)	LR (%)	RR (%)	LASSO (%)	Elastic net (%)	Ctree (%)	RF (%)	XGBoost (%)	Nnet (%)
5	31.9	25.8	25.1	24.7	Napp	28.4	28.7	30.6
10	40.6	39	40.1	40.1	33.0	41.7	41.8	41.7

Table 7. Avoided biopsies for a rate of CsPcCa wrongly classified (false negative rate). *Napp* not applicable, *LR* logistic regression, *RR* ridge regression, *Ctree* classification tree, *RF* random forest, *XGBoost* extreme gradient boosting, *Nnet* neural network.

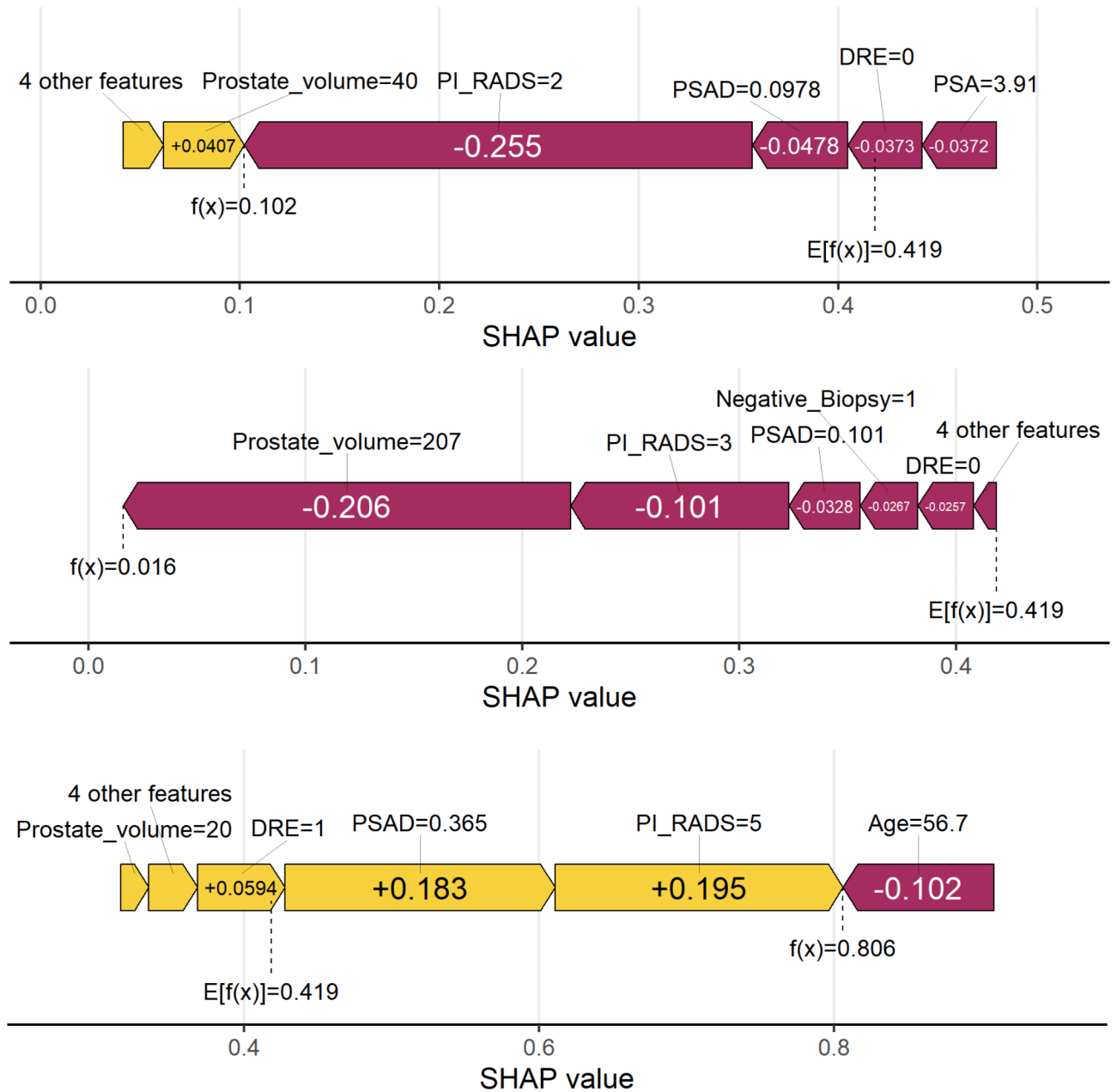


Fig. 13. SHAP values for 3 different cases (PI-RADS=2, 3, 5).

Data availability

The data presented in this study are available on request from the corresponding author.

Received: 22 September 2024; Accepted: 28 January 2025

Published online: 04 February 2025

References

1. Beam, A. L. et al. Artificial intelligence in medicine. *N. Engl. J. Med.* **388** (13), 1220–1221. <https://doi.org/10.1056/NEJMe2206291> (2023).
2. Michalski, R. S. & Carbonell, J. G. *T.M. Mitchell, Machine Learning: An Artificial Intelligence Approach* (Springer, 2013).
3. Sung, H. et al. Bray, global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71** (3), 209–249. <https://doi.org/10.3322/caac.21660> (2021).
4. Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B. & Srigley, J. R. P.A. Humphrey, grading committee, the 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *Am. J. Surg. Pathol.* <https://doi.org/10.1097/PAS.0000000000000530> (2016).
5. Van Poppel, H. et al. Serum PSA-based early detection of prostate cancer in Europe and globally: past, present and future. *Nat. Rev. Urol.* **19**, 562–572. <https://doi.org/10.1038/s41585-022-00638-6> (2022).

6. Louie, K. S., Seigneurin, A., Cathcart, P. & Sasieni, P. Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis. *Ann. Oncol.* **26** (5), 848–864. <https://doi.org/10.1093/annonc/mdu525> (2015).
7. Ahmed, H. U. et al. PROMIS study group, diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389** (10071), 815–822. [https://doi.org/10.1016/S0140-6736\(16\)32401-1](https://doi.org/10.1016/S0140-6736(16)32401-1) (2017).
8. Antolin, A. et al. The role of radiomics in the prediction of clinically significant prostate cancer in the PI-RADS v2 and v2.1 era: a systematic review. *Cancers (Basel)* **16** (17), 2951. <https://doi.org/10.3390/cancers16172951> (2024).
9. Lynch, C. J. & Liston, C. New machine-learning technologies for computer-aided diagnosis. *Nat. Med.* **24**, 1304–1305. <https://doi.org/10.1038/s41591-018-0178-4> (2018).
10. Azañe, F. Artificial intelligence for precision oncology: beyond patient stratification. *Npj Precis. Onc.* **3**, 6. <https://doi.org/10.1038/s41698-019-0078-1> (2019).
11. Haibe-Kains, B. et al. Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16. <https://doi.org/10.1038/s41586-020-2766-y> (2020).
12. Triquell, M. et al. Magnetic resonance imaging-based predictive models for clinically significant prostate cancer: a systematic review. *Cancers (Basel)* **14** (19), 4747. <https://doi.org/10.3390/cancers14194747> (2022).
13. Alberts, A. R. et al. Prediction of high-grade prostate cancer following multiparametric magnetic resonance imaging: improving the Rotterdam European randomized study of screening for prostate cancer risk calculators. *Eur. Urol.* **75**, 310–318. <https://doi.org/10.1016/j.eururo.2018.07.031> (2019).
14. Morote, J. et al. The Barcelona predictive model of clinically significant prostate cancer. *Cancers (Basel)* **14**, 1589. <https://doi.org/10.3390/cancers14061589> (2022).
15. Morote, J. et al. Comparison of Rotterdam and Barcelona magnetic resonance imaging risk calculators for predicting clinically significant prostate cancer. *Eur. Urol. Open Sci.* **53**, 46–54. <https://doi.org/10.1016/j.euro.2023.03.013> (2023).
16. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* (Springer, 2013).
17. Breiman, L., Friedman, J. H., Olshen, R. & Stone, C. J. *Classification and Regression Trees* (Chapman and Hall/CRC, 1984).
18. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
19. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29** (5), 1189–1232. <https://doi.org/10.1214/aos/1013203451> (2001).
20. Ghorbani, A. & Zou, J. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research* 97 2242–2251. <https://proceedings.mlr.press/v97/ghorbani19c.html> (2019).
21. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44** (3), 837–845 (1988).
22. Borque-Fernando, Á. et al. Gil-Sanz, 4Kscore test, prostate cancer prevention trial-risk calculator y European research screening prostate-risk calculator en la predicción del cáncer de próstata de alto grado; estudio preliminar. *Actas Urológicas Españolas* **40** (3), 155–163. <https://doi.org/10.1016/j.acuro.2015.09.006> (2016).
23. R Core Team. R: A language and environment for statistical computing. *R Found. Stat. Comput.* <https://www.R-project.org/> (2023).
24. Uddin, S., Khan, A., Hossain, M. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inf. Decis. Mak.* **19**, 281. <https://doi.org/10.1186/s12911-019-1004-8> (2019).
25. Ahmed, Z., Mohamed, K., Zeeshan, S. & Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database (Oxford)* **2020**, baaa010. <https://doi.org/10.1093/database/baaa010> (2020).
26. Hassija, V. et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* **16**, 45–74. <https://doi.org/10.1007/s12559-023-10179-8> (2024).
27. Hiremath, A. et al. An integrated nomogram combining deep learning, prostate imaging-reporting and data system (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. *Lancet Digit. Health* **3** (7), e445–e454. [https://doi.org/10.1016/S2589-7500\(21\)00082-0](https://doi.org/10.1016/S2589-7500(21)00082-0) (2021).
28. Wang, Z. et al. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. *IEEE Trans. Med. Imaging* **37** (5), 1127–1139. <https://doi.org/10.1109/TMI.2017.2789181> (2018).
29. Peng, T. et al. Can machine learning-based analysis of multiparameter MRI and clinical parameters improve the performance of clinically significant prostate cancer diagnosis? *Int. J. Comput. Assist. Radiol. Surg.* **16** (12), 2235–2249. <https://doi.org/10.1007/s11548-021-02507-w> (2021).
30. Donisi, L. et al. A combined radiomics and machine learning approach to distinguish clinically significant prostate lesions on a publicly available MRI dataset. *J. Imaging* **7** (10), 215. <https://doi.org/10.3390/jimaging7100215> (2021).
31. Ferro, M. et al. Radiomics in prostate cancer: an up-to-date review. *Ther. Adv. Urol.* **14**, 17562872221109020 (2022).
32. Mongan, J., Moy, L. & Kahn, C. E. Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).
33. Dominguez, I. et al. MRI-based surrogate imaging markers of aggressiveness in prostate cancer: development of a machine learning model based on radiomic features. *Diagnostics* **13**, 2779 (2023).
34. Morote, J. et al. Reducing the demand for magnetic resonance imaging scans and prostate biopsies during the early detection of clinically significant prostate cancer: applying the barcelona risk-stratified pathway in Catalonia. *Urol. Oncol.* <https://doi.org/10.1016/j.urolonc.2023.09.020> (2024).
35. Loeb, S. et al. Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.* **65** (6), 1046–1055. <https://doi.org/10.1016/j.eururo.2013.12.062> (2014).
36. Nordström, T. et al. Repeated prostate cancer screening using prostate-specific antigen testing and magnetic resonance imaging: a secondary analysis of the STHLM3-MRI randomized clinical trial. *JAMA Netw. Open* **7** (2), e2354577. <https://doi.org/10.1001/jamanetworkopen.2023.54577> (2024).
37. Bratt, O. Screening for prostate cancer with modern diagnostics—Another piece of the puzzle. *JAMA Netw. Open* **7** (2), e2354544. <https://doi.org/10.1001/jamanetworkopen.2023.54544> (2024).
38. Kattan, M. W. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. *Curr. Opin. Urol.* **13** (2), 111–116. <https://doi.org/10.1097/00042307-200303000-00005> (2003).
39. Shapley, L. S. Notes on the n-person game—ii: The value of an n-person game (1951).
40. Lin, Y. et al. Deep learning-based image quality assessment: impact on detection accuracy of prostate cancer extraprostatic extension on MRI. *Abdom. Radiol.* **49**, 2891–2901. <https://doi.org/10.1007/s00261-024-04468-5> (2024).
41. Li, H. et al. Machine learning in prostate MRI for prostate cancer: current status and future opportunities. *Diagnostics (Basel)* **12** (2), 289. <https://doi.org/10.3390/diagnostics12020289> (2022).
42. Mata, L. A., Retamero, J. A., Gupta, R. T., García Figueras, R. & Luna, A. Artificial intelligence-assisted prostate cancer diagnosis: radiologic-pathologic correlation. *Radiographics* **41** (6), 1676–1697. <https://doi.org/10.1148/rg.2021210020> (2021).
43. Mollura, D. J. et al. Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology* **297** (3), 513–520. <https://doi.org/10.1148/radiol.2020201434> (2020).
44. Du, X. et al. Effectiveness and cost-effectiveness of artificial intelligence-assisted pathology for prostate cancer diagnosis in Sweden: a microsimulation study. *Eur. Urol. Oncol.* **23** (24), S2588–9311. <https://doi.org/10.1016/j.euo.2024.05.004> (2024).

45. Sendra-Balcells, C. et al. Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. *Sci. Rep.* **13** (1), 2728. <https://doi.org/10.1038/s41598-023-29490-3> (2023).
46. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *Npj Digit. Med.* **1**, 18. <https://doi.org/10.1038/s41746-018-0029-1> (2018).
47. Wubineh, B. Z., Deriba, F. G. & Woldeyohannis, M. M. Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: a systematic literature review. *Urol. Oncol. Semin. Original Investig.* **42** (3), 48–56. <https://doi.org/10.1016/j.urolonc.2023.11.019> (2024).
48. Schoots, I. G. & Padhani, A. R. Risk-adapted biopsy decision based on prostate magnetic resonance imaging and prostate-specific antigen density for enhanced biopsy avoidance in first prostate cancer diagnostic evaluation. *BJU Int.* **127** (2), 175–178. <https://doi.org/10.1111/bju.15277> (2021).

Acknowledgements

L.M.E., M.E.E., J.E.-E. and A.B.-F. received support by the Government of Aragon [Grant Number T69_23R]; L.M.E. and A.B.-F. received support by the Ministerio de Ciencia e Innovación [Grant Number PID2020-116873GB-I00]; J.M.A., P.S. and J.M. received support by the Instituto de Salud Carlos III and the European Union [Grant Number PI20/01666].

Author contributions

Conceptualization: L.M.E., A.B.-F., J.M.; Data curation: L.M.E., J.M., J.M.A., P.S.; Formal analysis: L.M.E., M.E.E., J.E.-E.; Funding acquisition: L.M.E., J.M.; Investigation: A.B.-F., J.M.; Methodology: L.M.E., M.E.E., J.E.-E.; Software: L.M.E.; Supervision: A.B.-F., J.M.; Writing—original draft: L.M.E., J.M.; and Writing—review & editing: L.M.E., A.B.-F., M.E.E., J.E.-E., J.M.A., P.S., J.M.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.M.E.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025