3-17-2022

Machine learning-based prediction of upgrading on magnetic resonance imaging targeted biopsy in patients eligible for active surveillance

Bashier ElKarami

Mustafa Deebajah

Seth Polk

James O. Peabody

Behnam Shahrrava

See next page for additional authors

Authors
Bashier ElKarami, Mustafa Deebajah, Seth Polk, James O. Peabody, Behnam Shahrrava, Mani Menon, Abedalrhman Alkhateeb, and Shaheen Alanee

ELSEVIER

Urologic Oncology: Seminars and Original Investigations 000 (2022) 1−6

UROLOGIC ONCOLOGY

Clinical-Prostate cancer

Machine learning-based prediction of upgrading on magnetic resonance imaging targeted biopsy in patients eligible for active surveillance

Bashier ElKarami, Ph.D.[a], Mustafa Deebajah, M.D.[c,d], Seth Polk, M.S.[b], James Peabody, M.D.[c,d], Behnam Shahrrava, Ph.D.[a], Mani Menon, M.D.[c,d], Abedalrhman Alkhateeb, Ph.D.[a], Shaheen Alanee, M.D., M.P.H., M.B.A., F.A.C.S.[b,*]

[a] *Computer Science Department, The University of Windsor, ON, CA*
[b] *Department of Urology, Detroit Medical Center, Detroit, MI*
[c] *Department of Urology, Henry Ford Hospital, Detroit, MI*
[d] *Vattikuti Urology Institute, Detroit, MI*

Received 4 October 2021; received in revised form 20 December 2021; accepted 23 January 2022

Abstract

Objective: To examine the ability of machine learning methods to predict upgrading of Gleason score on confirmatory magnetic resonance imaging-guided targeted biopsy (MRI-TB) of the prostate in candidates for active surveillance.

Subjects and methods: Our database included 592 patients who received prostate multiparametric magnetic resonance imaging in the evaluation for active surveillance. Upgrading to significant prostate cancer on MRI-TB was defined as upgrading to G 3+4 (definition 1 - DF1) and 4+3 (DF2). Machine learning classifiers were applied on both classification problems DF1 and DF2.

Results: Univariate analysis showed that older age and the number of positive cores on pre-MRI-TB were positively correlated with upgrading by DF1 (P-value ≤ 0.05). Upgrading by DF2 was positively correlated with age and the number of positive cores and negatively correlated with body mass index. For upgrading prediction, the AdaBoost model was highly predictive of upgrading by DF1 (AUC 0.952), while for prediction of upgrading by DF2, the Random Forest model had a lower but excellent prediction performance (AUC 0.947).

Conclusion: We show that machine learning has the potential to be integrated in future diagnostic assessments for patients eligible for AS. Training our models on larger multi-institutional databases is needed to confirm our results and improve the accuracy of these models' prediction. © 2022 Elsevier Inc. All rights reserved.

Keywords: Machine learning; Targeted biopsy; Model; Prostate cancer

1. Introduction

The American Cancer Society estimates that, in 2021, there will be 248,530 new prostate cancer (CaP) cases diagnosed and 34,130 CaP deaths in the United States [1]. Active surveillance (AS) represented a paradigm shift in the management of CaP. The use of AS increased from 22% in 2004 to 2005 to 50% in 2014 to 2015 for patients with a Gleason score of 6 or below and from 9% in 2004 to 2005 to 13% in 2014 to 2015 for patients with a Gleason score of 7 or above [2]. Radiation therapy and surgery are 2 other options for treating PCa, but they are both associated with long lasting adverse effects on patients' urinary and sexual quality of life [3]. It has been shown that AS maintains excellent quality of life for PCa patients and is associated with excellent overall survival in well-designed prospective studies [4].

Recognizing the clinical need for accurate characterization of PCa in patients interested in AS, the scientific community has developed multiple biomarkers and imaging technologies to predict the upgrading of CaP diagnosed on prostate biopsy on final post-prostatectomy pathology. For this purpose, magnetic resonance imaging (MRI) and MRI

Source of funding: None.

*Corresponding author. Tel.: +1-313-745-8663; fax: +313-745-8222.
E-mail address: shaheen.alanee@gmail.com (S. Alanee).

https://doi.org/10.1016/j.urolonc.2022.01.012
1078-1439/© 2022 Elsevier Inc. All rights reserved.

based targeted biopsy (MRI-TB) emerged as 1 method of accurately characterizing the grade of CaP in patients who are candidates for AS [5]. However, MRI-TB of the prostate is an invasive procedure, and MRI of the prostate may not be available to some patients for reasons of comorbidities, patient weight restrictions, claustrophobia, and limited access [6,7]. Here, we report a machine learning-based model solution to this challenge. In this model, we used disease characteristics identified from systematic ultrasound-guided prostate biopsy and patient clinical features to predict Gleason score upgrading on MRI-TB of the prostate that can be directly used to guide treatment in clinical settings. This model, if validated, can be used in patients unfit, unwilling, or unable to access MRI-TB to counsel them on the safety of active surveillance of PCa.

2. Subjects and methods

After institutional review board approval, we performed a retrospective review of MRI-TB performed in patients managed with AS from October 2015 to February 2018 at a large health care organization. All MRI-transrectal ultrasound fusion-guided biopsies were performed by a single urologist with several years of experience in MRI-transrectal ultrasound targeted biopsy using the DynaCAD MRI-transrectal ultrasound fusion biopsy system UroNav (Invivo, Gainseville, FL). The biopsies were conducted using local anesthetic in the outpatient setting. Three target biopsy cores were taken from each lesion and were immediately followed by a 12-core systematic biopsy where 1 core was taken from each of the 12 sectors. Biopsies were scored by a fellowship-trained genitourinary pathologist.

Our database of 592 patients contained BMI, ethnicity, age, digital rectal examination results, PSA density (PSAD), number of pre-MRI biopsies, and number of positive cores, and results of MRI targeted biopsy. The workflow of our model is illustrated in Fig. 1. For this research, we applied our model on upgrading definition 1 (DF1) and upgrading definition 2 (DF2). DF1 was the detection of PCa GG > 3+3 on MRI-TB. DF2 was the detection of PCa GG > 3+4 on MRI-TB. We imputed the missing values with the average of the corresponding feature for each class. Initially, labeling the samples by upgrading vs. non-upgrading on MRI-TB led to a non-balance class prediction model biased toward the majority class. Therefore, preprocessing techniques were utilized for both datasets to handle the class imbalance issue and achieve higher prediction accuracy. For the first prediction model (DF1 vs. non-DF1), 2 techniques were applied; the synthetic minority over-sampling technique (SMOTE) was applied first to generate synthetic samples from the distribution of the minority class. Then, a simple random sampling with replacement was applied to produce a new subset containing equal DF1 and non-DF1 [8]. For the second prediction model (DF2 vs. non-DF2), SMOTE over-sampling technique is not enough due to the big difference between the number of samples between the majority and the minority classes. So both over-sampling and under-sampling techniques were used by selecting balanced ensembles with random samples from both classes. The classifier runs on the balanced ensembles that have the same number of samples in each class [9].

Overfitting is the main challenge in training the machine learning model. The model is trained on 90% of the data and tested on the remaining 10% each time to overcome overfitting. Then this process is repeated another 9 times, where the model test on another 10% each time. In the end, the model will be testing on the whole samples (100%) without using them in training (at each step ''Fold''). We tested different types of classifiers (Support Vector Machine, NaiveBayes, RandomForest, Bagging, and AdaBoostM1) via the standard 10-fold cross-validation experiments. An exhaustive search determined the parameters such as iteration number, base classifier, weight threshold, maximum depth of tree, and number of features. AdaBoost classifier and random forest classifiers achieved the best results DF1 dataset and DF2 dataset, respectively. In the first prediction model, AdaBoost classifier was used to classify DF1 vs. non-DF1 [10]. The balanced dataset was classified using 10-fold cross-validation; at each time in the loop, the AdaBoost classifier will take different samples from the one which is used in the previous steps. AdaBoost algorithm is a Boosting technique used as an Ensemble Method that combines multiple weak learners into a strong one. Weak learners are trained sequentially with weighted instances, where weights are updated and re-assigned each iteration, so higher weights are applied to incorrectly classified instances. AdaBoost $H(x)$ can be denoted as the following:

$$H(x) = sign\left(\sum_{t=1}^{T} \propto_t h_t(x)\right)$$

$$D_1(i) = 1/m$$

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\propto_t y_i h_t(x_i))}{Z_t}$$

Where:

x_i and y_i represent the instances and their labels.
$D_1(i)$ is the initial weights
$D_{t+1}(i)$ is the updated weights.
Z_t is a normalization factor.
$\propto_t$ is the weight updating parameter.

The second prediction model DF2 was built using a random forest classifier [11]. DF2 samples were tripled, and then 3 different ensembles were made by dividing the non-DF2 samples into 3 equal ensembles with different samples. Each ensemble of the non-DF2 matches the DF2 number of samples to have 3 balanced classification models. In each model we use 10-fold cross-validation with a random forest classifier to predict DF2 vs. non-DF2 samples.

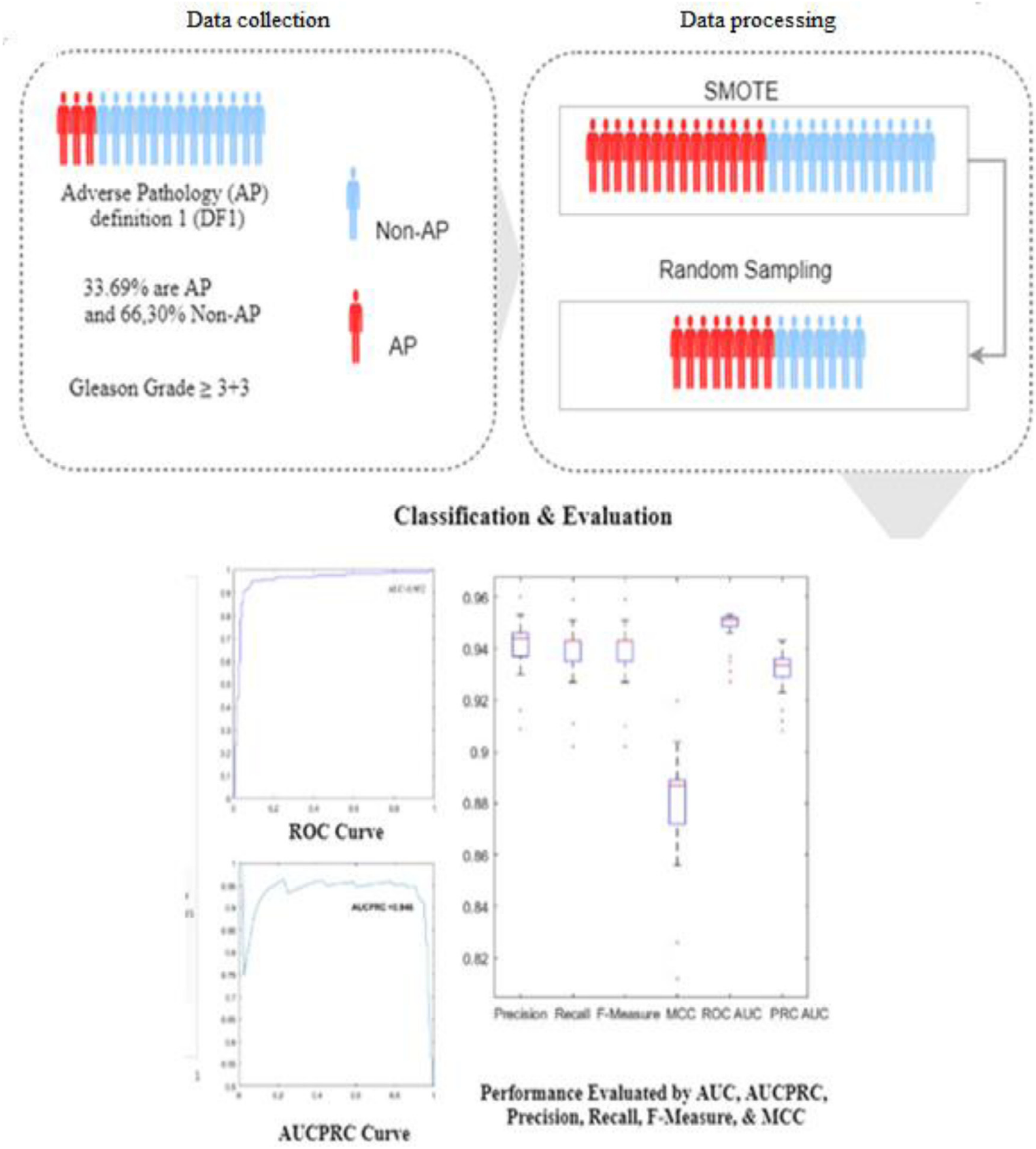

Fig. 1. The workflow of the machine learning model to predict upgrading of Gleason score on confirmatory magnetic resonance imaging guided targeted biopsy (MRI-TB) of the prostate in candidates for active surveillance.

Random forest prioritizes the features based on the GINI-index values of the features, where the 1 of higher value is the root of the tree and the next levels are for the less value, down to the least important features at the bottom of the tree. GINI-Index values are denoted as the following equation:$\mathrm{GINI} - \mathrm{Index} \;= 1 - \sum_{i=1}^{c} (p_i)^2$

Where p_i represents the frequency of the samples in class i, c represents the number of classes. Over/random sampling was applied to the whole dataset before we moved to the classification step.

The prediction performance was evaluated by Receiver Operating Characteristics curve (ROC), Accuracy, Precision, Recall, F-Measure, and Matthews Correlation coefficient. [12] We also provided the GINI-index values for clinical features to help readers verify that feature of importance in other types of models (logistic regression) are as well important in our model.

2.1. MRI acquisition and interpretation

All patients were imaged on 1 of 2 MRI systems: a GE Discovery 750 3.0-Tesla (GE Health care, Waukesha, WI), using a 32-channel torso phased array coil or a Philips Ingenia 3.0-Tesla (Philips Health care, Best, the Netherlands) using a 32-element anterior torso phased array coil coupled with an integrated posterior 20 element array in the tabletop.

All patients underwent a highly similar imaging protocol consisting of: large field of view (FOV) (32 cm or greater) 2-dimensional fast spin-echo T2-weighted images with fat suppression, and 3-dimensional T1 gradient-echo with Dixon fat-water separation (fat, water, in-phase, and out-of-phase reconstructions); small FOV (18 cm) fast spin-echo T2 images of the prostate in the axial, sagittal and coronal planes; axial diffusion weighted images in small FOV (Philips, 18 cm) and larger FOV (GE, 30 cm); small (22 cm) FOV bolus intravenous gadolinium chelate dynamic contrast enhanced T1 gradient-echo series (20 serial post-contrast phases, temporal resolution <10 seconds); and a final large FOV pelvic post-contrast T1 gradient-echo Dixon (water reconstruction) series. These studies were done without endorectal coil. Examinations were interpreted and analyzed using DynaCAD (InVivo, Gainesville, FL) by a single radiology group and followed PI-RADS v2 algorithms.

3. Results

Of patients who underwent MRI-TB ($n = 592$), 33.6% were upgraded by DF1 and DF2 upgraded 9.7%. Univariate analysis showed that older age and the number of positive cores on pre-MRI biopsy were positively correlated with upgrade by DF1 (P-value ≤ 0.05). Upgrade by DF2 was positively correlated with age and number of positive cores and negatively correlated with BMI. Baseline data used in Machine learning-based prediction on targeted biopsy in patients eligible for active surveillance are shown in Table 1.

Our model has achieved high performance in each statistical measurement including accuracy of 94.3% and 88.1%, precision of 94.6% and 88.0%, and recall of 94.3% and 88.1% for the DF1 (AdaBoost) and DF2 (random forest) respectively. The Area Under the Receiver Operating Characteristics Curve (AUC) was also high (95.2% for DF1 and 94.7% for DF2). The AUC and Area Under the Precision-Recall Curve were calculated and plotted to utilize their validity in an unbalanced dataset for further evaluation. Fig. 1 illustrates the evaluation of DF1 model performance.

Table 1
Baseline data used in Machine learning-based prediction on targeted biopsy in patients eligible for active surveillance.

	Values
Mean ± SD Age (RANGE)	65.0 ± 7.0 (49−83)
Mean ± SD (ng/mL) PSA (RANGE)	7.8 ± 6.0 (0.6−61.3)
Mean ± SD (mL) prostate volume (RANGE)	57.5 ± 33.4 (15−195)
Mean ± SD (months) Interval from previous biopsy (RANGE)	28.7 ± 15.2 (1−132)
Ethnicity = African American (%)	24.20%
Ethnicity = White (%)	60.00%
Ethnicity = Other (%)	15.80%
Percent Abnormal DRE	30.60%
Final Pathology = Benign (%)	24.30%
Final Pathology = 3 + 3 (%)	42.50%
Final Pathology = 3 + 4 (%)	23.90%
Final Pathology = 4 + 3 (%)	8.60%
Final Pathology = 4 + 4 (%)	0.74%
Mean ± SD Number of positive cores on immediate pre-MRI-TB systematic biopsy (range)	1.9 ± 2.9 (0−6)

To determine which clinical feature plays a crucial role in predicting upgrading of Gleason score, we calculated the feature importance through GINI-Index values in random forest. We found that PSAD is the most important, and digital rectal examination (DRE) is the least important, as shown in Fig. 2.

To thoroughly evaluate each feature's effect on the model's accuracy, we also run ablation tests (test different combinations of features, e.g., leave 1 out in each test). The PSA density feature came up as the most critical feature. The Age feature and BMI showed a small difference between them. The rest of the features had little importance compared to the first 3 features (Table 2).

4. Discussion

We found that machine learning provided highly predictive models of upgrading after MRI-TB, which our ROC of >94% reflected. Machine learning methods are good at constructing models on a complex or non-linear dataset, and random forest applies the strategy of training on different parts of the dataset and averaging multiple decision trees to reduce the variance and avoid overfitting; thus, tree-based methods, especially random forest, are more suitable for such problems.[11] Our work provides an attempt to integrate cost-effective clinical variables with new analysis methods to answer clinical questions instead of employing more expensive predictors such as advanced imaging or biomarkers.

There have been many attempts at developing biomarkers to select patients for active surveillance; however, in many cases, the application of the biomarker in clinical practice was limited by considerations of cost, feasibility, and accuracy. One such biomarker is the 17-gene Oncotype DX Genomic Prostate Score (GPS) test that was promoted as a predictor of upgrading in patients with low-risk CaP treated with immediate surgery. Lin et al. evaluated the GPS test as a predictor of outcomes in a multicenter active surveillance cohort where diagnostic biopsy tissue was obtained from men enrolled at 8 sites in the Canary Prostate Active Surveillance Study [13]. The primary endpoint was adverse pathology (AP) (Gleason Grade Group [GG] ≥3, ≥pT3a) in men who underwent radical prostatectomy after initial surveillance. The GPS results were obtained for 432 men (median follow-up, 4.6 years). GPS was not predictive of AP when the authors adjusted for prostate-specific antigen density, and no association was observed between GPS and subsequent biopsy upgrade ($P = 0.48$). The authors concluded that adding GPS to a model containing PSA density and diagnostic Gleason grade did not significantly improve

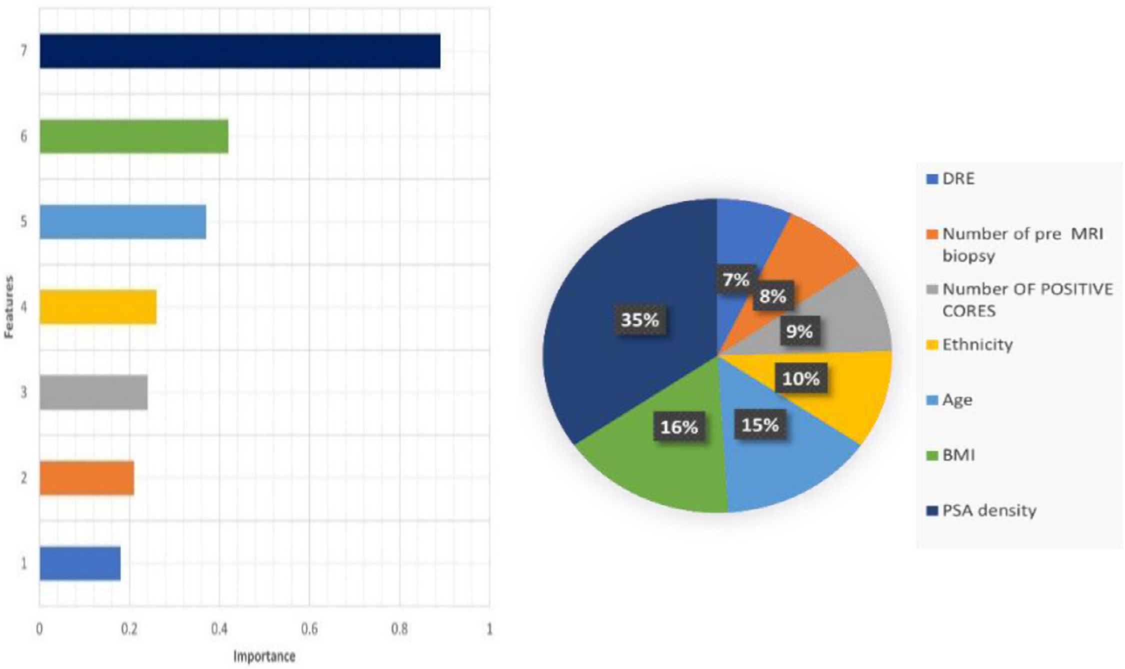

Fig. 2. The importance of the clinical features calculated using random forest.

stratification of risk for AP over the clinical variables alone. We used diagnostic Gleason grade and PSA density in our modeling and used them to predict upgrading on biopsy accurately. Their results support clinical variables as effective predictors of upgrading and a good proxy of biology at the molecular level.

Along the same lines, An American Society of Clinical Oncology multidisciplinary expert panel, with representatives from the European Association of Urology, American Urological Association, and the College of American Pathologists, conducted a systematic literature review of localized CaP biomarker studies between January 2013 and January 2019 [14]. Of 555 studies identified by the panel, 77 were selected for inclusion, plus 32 additional references selected by the Expert Panel. The panel highlighted Oncotype Dx Prostate, Prolaris, Decipher, Decipher PORTOS, and ProMark in their analysis and identified a paucity of prospective studies assessing short- and long-term outcomes of patients when these markers are integrated into clinical decision making. The panel then concluded that tissue-based molecular biomarkers are not recommended for routine use and that they may improve risk stratification only when added to standard clinical parameters. This panel's guidelines support our approach to retooling available clinical parameters using new analysis methods to predict upgrading. Nevertheless, genomic markers continue to have a yet to be explored potential for future integration in machine learning- based predictive models to further increase their predictive accuracy

Table 2
Results of the ablation tests to evaluate effects of different feature on machine-learning based prediction of adverse pathology on magnetic resonance imaging targeted biopsy of the prostate.

Features	Accuracy
All features	94.3
BMI Out	92.68
Age Out	88.61
PSA Out	87.8
DRE Out	94.3
Number of positive cores in immediate pre MRI-TB systematic biopsy Out	95.12
Number of pre- MRI-TB systematic biopsy Out	95.12

We used clinical variables and machine learning to predict upgrading on MRI-TB of the prostate in patients who are candidates for active surveillance. Through well designed research, MRI-targeted biopsy was found to detect significantly more grade progressions in active surveillance patients compared to systematic biopsy providing compelling evidence that prostate MRI and MRI-TB should be included in current active surveillance protocols [5]. However, MRI of the prostate is not accessible to many patients due to difficulty contacting patients and insurance denials, and African-American patients are disproportionately affected by barriers to MRI of the prostate in the course of Active surveillance [6,7]. Modeling clinical variables with advanced methods such as machine learning could allow us to use lessons learned from well-funded studies involving advanced imaging technologies to manage patients in resource-limited environments with limited access to these technologies.

This study has limitations. One limitation is missing values. To impute the missing values, we replaced them with the average of the corresponding feature for each class, which may introduce biases, and intensify the batch effects among cohorts and complexity among features. Having missing values in the dataset should be a strong reason why linear, logistic regression, and Cox regression fail to generate highly predictive models. A more intact dataset should generate more precise models. In addition, machine learning models mainly focus on prediction accuracy as a performance metric because it is sometime hard to explain the process of prediction. Therefore, machine learning, is commonly described as black-box models. To overcome this limitation in our analysis, provided the GINI-index values for clinical feature to enable clinicians to know what clinical feature is the model using to make predictions so as to give them insights into how this model is working and if the model is classifying patients into different classes (DF1 vs. DF2) for the right reasons.

Overall, our research shows that machine learning has the potential to be integrated in future diagnostic assessments for patients eligible for AS. Training our models on larger multi-institutional databases is needed to confirm our results and improve the accuracy of these models' prediction.

Conflicts of interest

The authors have no conflict of interest.

References

[1] American Cancer Society. Cancer facts & figures 2021. Atlanta: American Cancer Society; 2021 https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf.

[2] Liu Y, Hall IJ, Filson C, et al. Trends in the use of active surveillance and treatments in Medicare beneficiaries diagnosed with localized prostate cancer. Urol Oncol 2021;39(7):432.e1–10.

[3] Barocas DA, Alvarez J, Resnick MJ, et al. Association between radiation therapy, surgery, or observation for localized prostate cancer and patient-reported outcomes after 3 years. JAMA 2017;317(11):1126–40.

[4] Hamdy FC, Donovan JL, Lane JA, et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. N Engl J Med 2016;375(15):1415–24.

[5] Yerram NK, Long L, O'Connor LP, et al. Magnetic resonance imaging-targeted and systematic biopsy for detection of grade progression in patients on active surveillance for prostate cancer. J Urol 2020. https://doi.org/10.1097/JU.0000000000001547.

[6] Rosenkrantz AB, Lepor H, Huang WC, et al. Practical barriers to obtaining pre-biopsy prostate MRI: assessment in over 1,500 consecutive men undergoing prostate biopsy in a single urologic practice. Urol Int 2016;97(2):247–8.

[7] Walton EL, Deebajah M, Keeley J, et al. Barriers to obtaining prostate multiparametric magnetic resonance imaging in African-American men on active surveillance for prostate cancer. Cancer Med 2019;8(8):3659–65.

[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[9] Elkarami B, Alkhateeb A, Rueda L. Cost-sensitive classification on class-balanced ensembles for imbalanced non-coding RNA data. In: 2016 IEEE EMBS International Student Conference (ISC); 2016. p. IEEE1–4.

[10] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.

[11] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[12] Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manage. Proc 2015;5(2):1.

[13] Lin DW, Zheng Y, McKenney JK, et al. 17-tene genomic prostate score test results in the canary prostate active surveillance study (PASS) cohort. J Clin Oncol 2020;38(14):1549–57.

[14] Eggener SE, Rumble RB, Armstrong AJ, et al. Molecular biomarkers in localized prostate cancer: ASCO Guideline. J Clin Oncol 2020;38(13):1474–94.