Original Article

Machine Learning Approaches for the Prediction of Prostate Cancer according to Age and the Prostate-Specific Antigen Level

Jaegeun Lee[1], Seung Woo Yang[1], Seunghee Lee[2], Yun Kyong Hyon[2], Jinbum Kim[3], Long Jin[1], Ji Yong Lee[1], Jong Mok Park[1], Taeyoung Ha[2], Ju Hyun Shin[1], Jae Sung Lim[1], Yong Gil Na[1], Ki Hak Song[1]

[1]Department of Urology, Chungnam National University College of Medicine, Daejeon, Korea
[2]Division of Medical Mathematics, National Institute for Mathematical Sciences, Daejeon, Korea
[3]Department of Urology, Konyang University College of Medicine, Daejeon, Korea

Purpose: The aim of this study was to evaluate the applicability of machine learning methods that combine data on age and prostate-specific antigen (PSA) levels for predicting prostate cancer.
Materials and Methods: We analyzed 943 patients who underwent transrectal ultrasonography (TRUS)-guided prostate biopsy at Chungnam National University Hospital between 2014 and 2018 because of elevated PSA levels and/or abnormal digital rectal examination and/or TRUS findings. We retrospectively reviewed the patients' medical records, analyzed the prediction rate of prostate cancer, and identified 20 feature importances that could be compared with biopsy results using 5 different algorithms, viz., logistic regression (LR), support vector machine, random forest (RF), extreme gradient boosting, and light gradient boosting machine.
Results: Overall, the cancer detection rate was 41.8%. In patients younger than 75 years and with a PSA level less than 20 ng/mL, the best prediction model for prostate cancer detection was RF among the machine learning methods based on LR analysis. The PSA density was the highest scored feature importances in the same patient group.
Conclusions: These results suggest that the prediction rate of prostate cancer using machine learning methods not inferior to that using LR and that these methods may increase the detection rate for prostate cancer and reduce unnecessary prostate biopsy, as they take into consideration feature importances affecting the prediction rate for prostate cancer. **(Korean J Urol Oncol 2019;17:110-117)**

————————

Key Words: Prediction · Prostate cancer · Machine learning · Prostate biopsy

INTRODUCTION

Prostate cancer is the most common malignancy among men in the Organization for Economic Co-operation and Development countries and the second most common malignancy worldwide.[1,2] According to GLOBOCAN data, there were 1,276,106 new prostate cancer cases and 358,989 prostate-cancer-related deaths in 2018, and the number of newly diagnosed cases is expected to double by 2040 (2,293,818); however, mortality is predicted to increase by 1.05%.[3] In South Korea, the

incidence of prostate cancer continues to increase, but the recent mortality rate has decreased rapidly.[4] Since Billroth first performed radical perineal prostatectomy in 1867, the surgical technique has undergone several developments to robotic-assisted radical prostatectomy.[5] In addition to radiation therapy, which has improved survival rates through the development of equipment and combination with hormonal therapy, the development of new drugs, such as androgen signaling target drugs and chemotherapeutic agents, continues to reduce mortality rates associated with single and combined therapy. To facilitate early diagnosis and to avoid unnecessary prostate biopsies, multiparametric magnetic resonance imaging (mpMRI) using a Prostate Imaging-Reporting and Data System (PI-RADS)[6] and biomarkers such as free prostate-specific antigen (PSA), total/free PSA ratio and 4Kscore,[7] prostate health index,[8] and PCA3 (prostate cancer antigen 3)[9] may be used to diagnose prostate cancer. The developed techniques of surgery, medicine, and diagnostic tools could help to reduce mortality of prostate cancer.

Recently, artificial intelligence technology and machine learning methods have been applied to analyze large amounts of data the medical field, and their adequacy and usefulness in diagnosis are increasing.[10] Analysis using a combination of machine learning methods and mammograms has led to an accurate diagnosis of breast cancer,[11] and an automatic grading system has been applied to determine the Gleason grade of prostate cancer using histopathological images.[12] To our knowledge, there have been no studies to evaluate the prediction rate of prostate cancer using machine learning methods in Korea until now. Thus, in this study, we evaluated the applicability of machine learning methods that combine data on age and PSA levels in predicting prostate cancer.

MATERIALS AND METHODS

This study was approved by the Ethics Committee of Chungnam National University Hospital (#2018-12-055). We analyzed 1,300 patients who underwent transrectal ultrasonography (TRUS)-guided prostate biopsy at Chungnam National University Hospital between 2014 and 2018 because of elevated PSA levels and/or abnormal digital rectal examination and/or TRUS findings. Prostate biopsies were performed as outpatient procedures. All patients underwent a 12-core TRUS-guided conventional systematic biopsy with the 18-gauge biopsy needle using Flex Focus 400 ultrasound equipment with a 5–10 MHz bi-convex probe (BK medical, Peabody, MA, USA).

We retrospectively reviewed the patients' medical records for underlying diseases such as diabetes, hypertension, and cardiovascular disease as well as their laboratory findings such as white blood cell (WBC) count and hemoglobin, albumin, alkaline phosphatase, and serum PSA levels, and 943 patients were included in the study on the basis of feature importances and analytic factors. On the basis of pathologic confirmation, the patients were divided into 2 groups, namely, the prostate cancer group (n=394) and the nonprostate cancer group (n=549), which had benign prostate tissue, prostate hypertrophy, inflammation of the prostate, prostate atrophy, and atypical cells, etc.

Thereafter, we analyzed the prediction rate of prostate cancer and identified 20 feature importances that could be compared with biopsy results using 5 different algorithms: logistic regression (LR),[13] support vector machine (SVM),[14] random forest (RF),[15] extreme gradient boosting (XGB),[16] and light gradient boosting machine (LGBM).[17] While classical LR is one of the most common utilized linear statistical models for discriminant analysis and for clinical classification and regression problems, we expect that these machine learning algorithms may perform better than LR, which is based on generalized linear regression. SVM is much more geometrically motivated by finding the optimal hyperplane between target classes. We expect SVM to perform marginally better than LR. Conversely, the tree-based machine learning algorithms are the state of the art for structured data. In this manner, we applied RF, XGB, and LGBM. Compared to other machine learning methods, tree-based methods provide interpretation of the results via feature importance scores, which is the advantage of these methods that is corroborated by our results. For all machine learning prediction models, 80% of the randomly chosen samples of data (same data set for each algorithms) were used for training, the remaining 20% were used as a test set. The predicting models were trained the data set combined with augmented samples generated by synthetic minority oversampling technique (SMOTE). Unlike random over sampling, SMOTE used interpolation to create new observations near existing observations of the minority class. To prevent outliers from shrinking the decision boundary between 2 classes, we used outlier rejection of K=2 nearest neighbors. The performance of the predictive

models was evaluated through an average of 20 samplings to prevent biased prediction and overfitting. When establishing the predictive model, typically, hyper-parameters for each algorithm were adopted for the classifications. The characteristics of the patients in the prostate cancer group and nonprostate cancer group were compared using Pearson chi-square test for categorical variables and Student t-test for continuous variables in IBM SPSS Statistics ver. 21.0 (IBM Co., Armonk, NY, USA).

RESULTS

Of the 943 patients, 394 (41.8%) were diagnosed with prostate cancer, and the mean age (69.7±6.8 years), WBC (6,899.3±1,834.1/μL), PSA (20.6±20.2 ng/mL), PSA density (0.64±0.66 ng/mL/cm^3) and PZPSAD (0.93±0.97 ng/mL/cm^3) of the prostate cancer group were significantly higher than those (mean age, 67.0±6.4 years; WBC, 6,650.7±1,785.8/μL; PSA, 8.8±8.1 ng/mL; PSA density, 0.38±0.32 ng/mL/cm^3; and PZPSAD, 0.38±0.32 ng/mL/cm^3) of the nonprostate cancer

Table 1. Clinical characteristics according to the results of pathological analysis

Characteristic	Pathologic results from prostate biopsy		p-value
	Nonprostate cancer (n=549)	Prostate cancer (n=394)	
Age (yr)*	67.0±6.4	69.7±6.8	<0.001
Diabetes mellitus[†]			0.560
No	454 (82.7)	320 (81.2)	
Yes	95 (17.3)	74 (18.8)	
Hypertension[†]			0.151
No	288 (52.5)	188 (47.7)	
Yes	261 (47.5)	206 (52.3)	
CVA history[†]			0.791
No	507 (92.3)	362 (91.9)	
Yes	42 (7.7)	32 (8.1)	
Pyuria[†]			0.758
No	497 (90.5)	359 (91.1)	
Yes	52 (9.5)	35 (8.9)	
Bacteriuria[†]			0.094
No	397 (72.3)	265 (67.3)	
Yes	152 (27.7)	129 (32.7)	
Hb (g/dL)*	14.5±1.2	14.2±1.5	0.011
WBC (/μL)*	6,650.7±1,785.8	6,899.3±1,834.1	0.037
GFR (mL/min/1.73 m^2)*	90.7±19.9	89.7±22.2	0.456
AST (U/L)*	21.9±8.3	22.2±10.3	0.617
ALT (U/L)*	21.7±12.7	20.4±11.1	0.125
ALP (U/L)*	70.8±19.4	73.7±52.0	0.292
Albumin (g/dL)*	4.2±0.3	4.1±0.3	0.002
PSA (ng/mL)*	9.9±8.1	20.6±20.2	<0.001
Prostate volume (cm^3)*	48.3±20.6	36.0±15.7	<0.001
Prostate transitional volume (cm^3)*	20.0±14.0	12.2±9.1	<0.001
Prostate peripheral volume (cm^3)*	28.3±9.5	23.7±8.6	<0.001
PSA density (ng/mL/cm^3)*	0.23±0.21	0.64±0.66	<0.001
PZPSAD (ng/mL/cm^3)*	0.38±0.32	0.93±0.97	<0.001
Biopsy count	1.3±0.6	1.1±0.4	<0.001

CVA: cardiovascular accidents, Hb: hemoglobin, WBC: white blood cell count, GFR: glomerular filter rate, AST: aspartate aminotransferase, ALT: alanine aminotransferase, ALP: alkaline phosphatase, PSA: prostate-specific antigen, PZPSAD: peripheral zone PSA density.
*Independent t-test. [†]Chi-square test or Fisher exact test.

group (p<0.05). Hemoglobin (14.2±1.5 g/dL), albumin (4.1±0.3 g/dL), prostate volume (36.0±15.7 cm^3), prostate transitional volume (12.2±9.1 cm^3), prostate peripheral volume (23.7±8.6 cm^3), and biopsy counts (1.1±0.4) in the prostate cancer group were significantly lower than those (hemoglobin, 14.5±1.2 g/dL; albumin, 4.2±0.3 g/dL; prostate volume, 48.3±20.6 cm^3; prostate transitional volume, 20.0±14.0 cm^3), prostate peripheral volume (28.3±9.5 cm^3), and biopsy counts

(1.3±0.6) of the nonprostate cancer group (p<0.05); however, there was no difference in the incidence of underlying diseases, such as diabetes, hypertension, and cardiovascular disease, between the 2 groups (Table 1).

The prediction rates for prostate cancer, according to age and PSA level, using various machine learning methods were as follows (median detection rate [minimum to maximum]): SVM, 69.2% (43.8%–78.3%); RF, 72.1% (65.0%–81.2%); XGB,
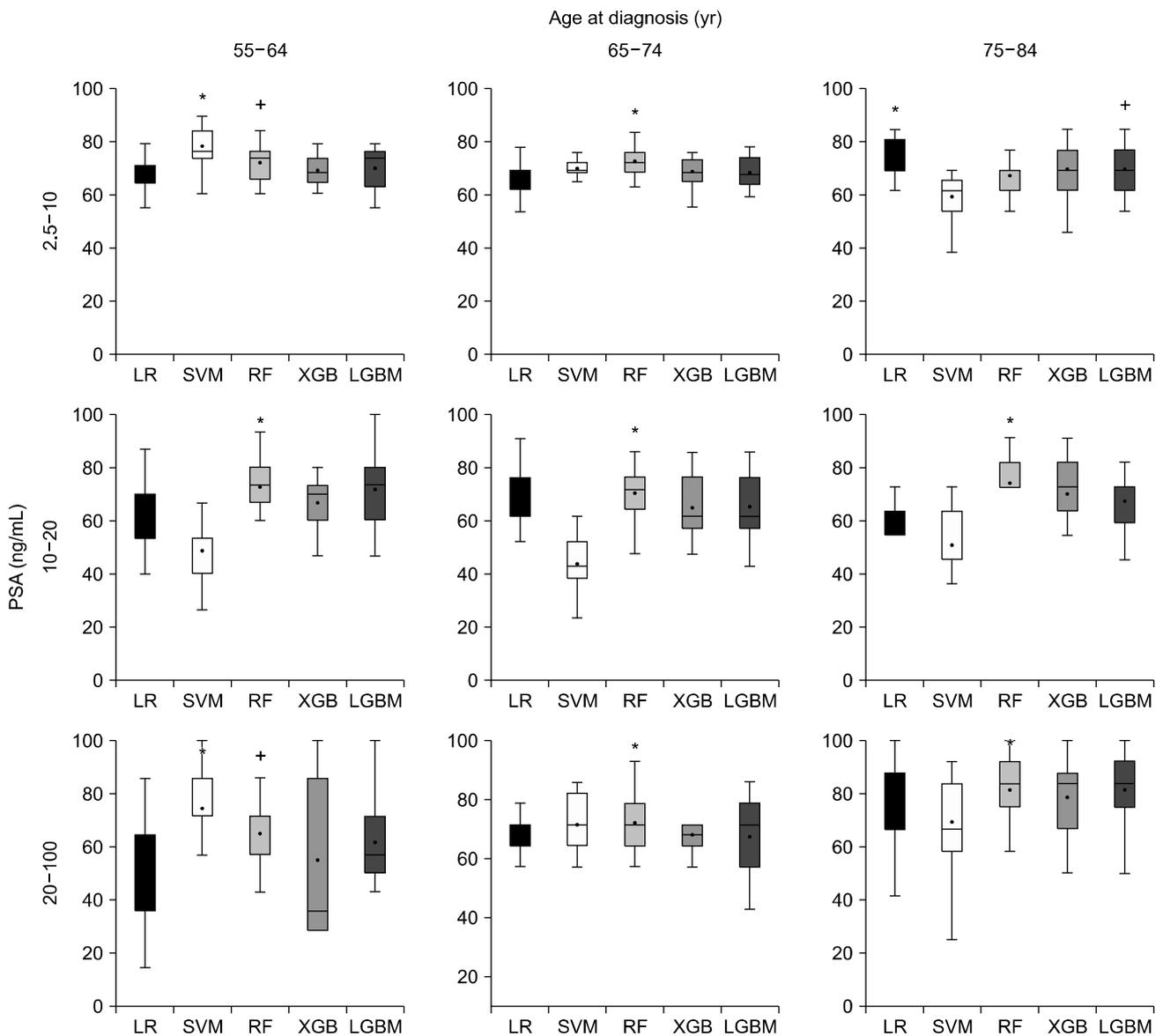

Fig. 1. Prediction rate of prostate cancer using machine learning methods. The RF method was the best method for prediction of prostate cancer, except in patients who were 55–64 years old or 75–84 years old and had a PSA level of 2.5–10 ng/mL and in patients who were 55–64 years old and had a PSA level of 20–100 ng/mL. PSA: prostate-specific antigen, LR: logistic regression, SVM: support vector machine, RF: random forest, XGB: extreme gradient boosting, LGBM: light gradient boosting machine. *Best prediction method for prostate cancer according to age and PSA level. [†]Better prediction method.

68.6% (55.0%−78.3%); LGBM, 68.3% (61.4%−81.2%); and LR, 67.1% (50.0%−75.4%). In 8 out of 9 groups, except in patients older than 75 years and with a PSA level of 2.5−10 ng/mL, the best prediction model for prostate cancer detection was the machine learning method. Among patients older than 75 years and with a PSA level less than 10 ng/mL, LR showed the best prediction rate (74.6%). In patients with a PSA level of 10−20 ng/mL or the 65- to 74-year age group, the best prediction model for prostate cancer detection was RF among the machine learning methods compared with LR analysis (Fig. 1).

The feature importances that affected the detection rate of prostate cancer were analyzed using the best machine learning method. SVM is a statistical-based machine learning method, and it cannot derive the feature importances influencing the de-
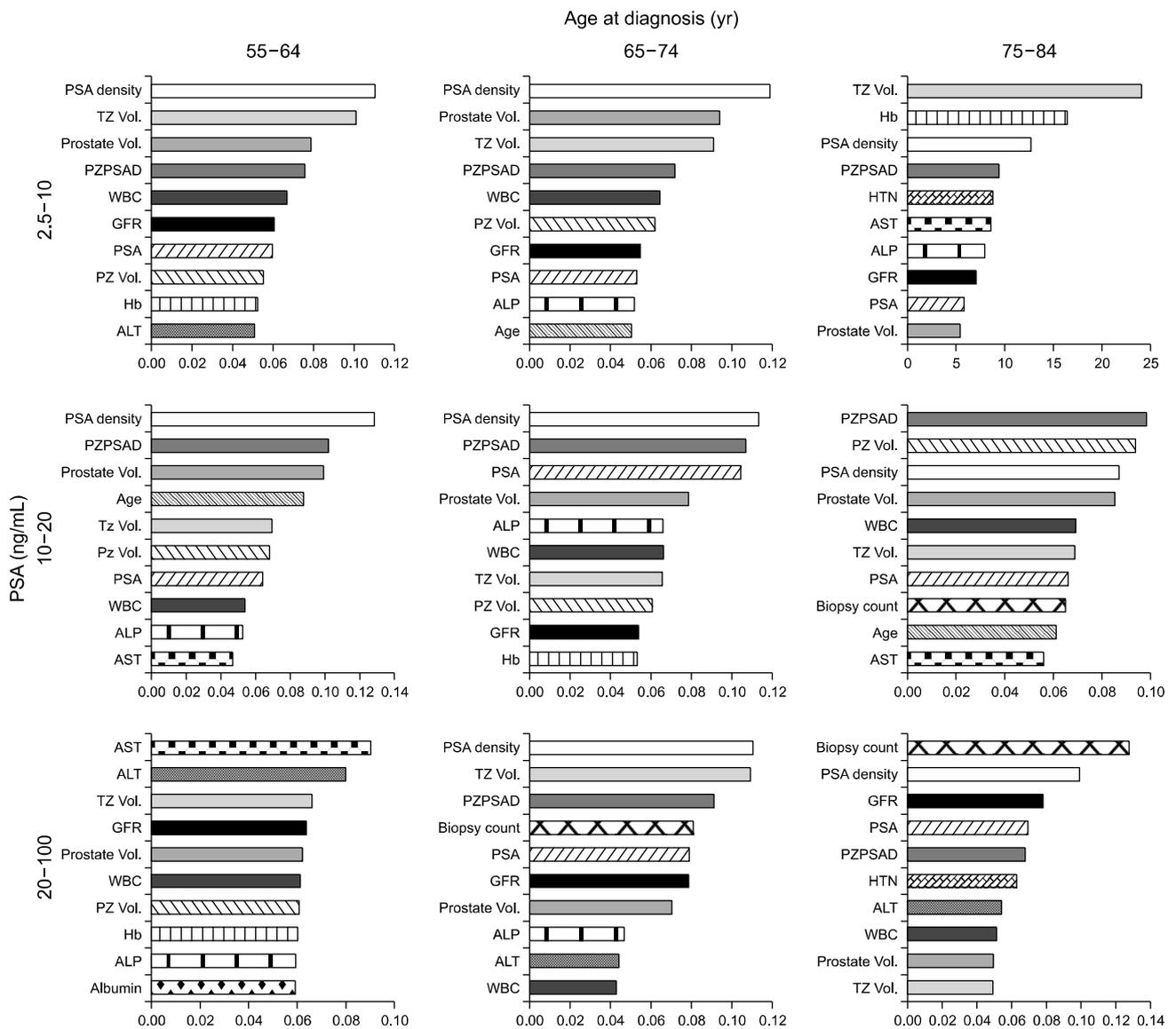

Fig. 2. Feature importance according to age and PSA level. PSA density was the highest scored feature importance in patients who were 55−74 years old and had a PSA level of 2.5−20 ng/mL. In patients who were older (75−84 years) and had a low PSA level (2.5−10 ng/mL), transitional zone volume influenced cancer prediction; previous biopsy history also affected cancer prediction in patients with a high PSA level (20−100 ng/mL) and of older age (75−84 years). PSA: prostate-specific antigen, TZ vol.: transitional zone volume, PZPSAD: peripheral zone PSA density, WBC: white blood cell, GFR: glomerular filter rate, PZ vol.: peripheral zone volume, Hb: hemoglobin, ALT: alanine aminotransferase, ALP: alkaline phosphatase, AST: aspartate aminotransferase.

tection rate analysis, so we used the 2nd ranked tree-based machine learning algorithms to identify feature importances. PSA density was highest scored feature importance in patients younger than 75 years and with a PSA level less than 20 ng/mL. Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels were identified as feature importances in patients with aged less than 65 years and with a PSA level of 20–100 ng/mL. However, in patients older than 75 years, the feature importances were different. Among them, transitional zone volume was the highest scored feature importance in patients with a PSA level less than 10.0 ng/mL (Fig. 2).

DISCUSSION

The incidence and mortality rates of prostate cancer were higher in Western society than in Eastern society.[18] Recently, it was reported that in South Korea, prostate cancer has a tendency to be expressed in patients in their 50s, similar to that in Western society.[4] Prostate cancer is the most frequent cancer in men, and most cases are diagnosed at an age of approximately 66 years. Of the patients diagnosed with prostate cancer, 69% die at an age older than 75 years.[19] Therefore, if prostate biopsy is performed considering the factors and feature importances that influence the detection rate of prostate cancer according to age, the detection rate of prostate cancer can be increased and the complications of prostate biopsy and unnecessary prostate biopsy, reduced, especially in elderly patients. Takeuchi et al.[20] reported that the prediction rate for prostate cancer improved by about 5%–10% when using a multilayer artificial neural network compared with LR analysis in 334 patients who underwent 3.0T mpMRI before TRUS-guided prostate biopsy. In this study, various machine learning methods were used to calculate the prediction rate for prostate cancer compared with conventional LR analysis. In patients aged over 75 years and with a PSA level of 2.5–10 ng/mL, LR analysis showed the best prediction rate (74.6%); however, machine learning methods were better than LR analysis in other patient groups. In particular, RF (prediction rate, 70.5%–72.4%) was the better than LR analysis (prediction rate, 65.6%–70.0%) in patients aged 65–74 years (Fig. 1). In South Korea, it is difficult to apply mpMRI owing to several cost issues and insurance coverage; however, we believe that it is best to apply mpMRI using the PI-RADS for the detection of prostate cancer. In this study, we also analyzed the feature importances influencing the prediction rate of prostate cancer in each patient group. PSA density was the highest scored feature importance in patients aged below 75 years and with a PSA level less than 20 ng/mL (Fig. 2). Several reports have shown that PSA density is one of the most important predictors of the detection of prostate cancer.[21,22] Sfoungaristos and Perimenis[23] reported that PSA density was a better predictor than the PSA level and Gleason score for poor pathologic prognostic factors such as a positive surgical margin, extracapsular disease, seminal vesicle involvement, and lymph node involvement. In our study, several feature importances that were less related to prostate cancer, such as AST and ALT levels and glomerular filter rate, were assessed in patients with aged below 65 years and with a PSA level of 20–100 ng/mL because of the small sample size of only 35 patients and the high detection rate of prostate cancer (26 patients, 74.3%) in patients in this group. However, in patients older than 75 years, the feature importances were different. Among them, transitional zone volume was the most feature importance in patients with a PSA level less than 10.0 ng/mL. It must be considered that benign prostatic hyperplasia influenced the elevation of PSA in this age group. In patients older than 75 years and a PSA level greater than 20.0 ng/mL, the most feature importance was a previous prostate biopsy history, and the detection rate for prostate cancer was 72.1%. It seems that the indication for biopsy was highly restricted, and biopsy was performed whenever necessary in this group. Hamilton et al.[24] studied the types of treatments chosen for patients with moderate-to-high-risk prostate cancer. Of the patients aged less than 75 years, 88.3% were treated with the aim of complete cure, but only 40.7% of the patients aged 75 years or older in the same risk group wanted to be treated with intent for cure. This implies that patients belonging to relatively older age groups choose curative treatment at a lower rate. In addition, the International Society of Geriatric Oncology recommended that patients aged 70 years or older with prostate cancer be managed with appropriate treatments after considering their health status, comorbidity, and cognitive function, and conservative or palliative managements be considered if the patients belonged to group 3 or 4 on geriatric screening using Geriatric 8 or mini-COG.[19] Therefore, in most patients with a PSA level greater than 10-ng/mL prostate biopsy should be performed. However, as is known, it is associated with a smaller benefit to surgical treatment over active surveillance in patients over 75 years of age and in those with poor health

status.[25] In terms of the minor benefits of cancer specific mortality and life expectancy, determining the feature importances that influence the prediction rate of prostate cancer may help to avoid unnecessary biopsy in patients of the older age groups, and it may be more persuasive for younger patient group.

We recognize that our present study has several limitations. This study is retrospective and evaluated outcomes for a relatively small number of patients. Additional randomized, prospective multicenter studies are needed to confirm our data. Furthermore, as medical records were sparse and the number of patients in each group were small, it was difficult to ensure sufficient training for machine learning. And the application of mpMRI using the PI-RADS may improve the prediction rate. Considering the low cancer-specific mortality in low-risk prostate cancer, further research is needed to predict clinically significant disease. Gleason score and TNM staging should be included in future studies for unfavorable prostate cancer.

CONCLUSIONS

In the future, using prostate MRI combined with artificial intelligence or a variety of newly developed biomarkers will markedly improve the diagnosis rate of prostate cancer and allow direct diagnosis of prostate cancer without a prostate biopsy. In the current scenario, this study will be useful in determining the requirement for prostate biopsy by considering the feature importances according to age and PSA levels for the prediction of prostate cancer.

CONFLICT OF INTEREST

The authors claim no conflicts of interest.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144:1941-53.
3. Rawla P. Epidemiology of prostate cancer. World J Oncol 2019;10:63-89.
4. National Cancer Center. 2016 Korea National Cancer registration. Goyang (Korea): National Cancer Center; 2019.
5. Hatzinger M, Hubmann R, Moll F, Sohn M. The history of prostate cancer from the beginning to DaVinci. Aktuelle Urol 2012;43:228-30.
6. Pickersgill NA, Vetter JM, Andriole GL, Shetty AS, Fowler KJ, Mintz AJ, et al. Accuracy and variability of prostate multiparametric magnetic resonance imaging interpretation using the prostate imaging reporting and data system: a blinded comparison of radiologists. Eur Urol Focus 2018 Oct 13 [Epub]. pii: S2405-4569(18)30301-8. https://doi.org/10.1016/j.euf.2018.10.008.
7. Lin DW, Newcomb LF, Brown MD, Sjoberg DD, Dong Y, Brooks JD, et al. Evaluating the four kallikrein panel of the 4Kscore for prediction of high-grade prostate cancer in men in the canary prostate active surveillance study. Eur Urol 2017;72:448-54.
8. Loeb S, Sanda MG, Broyles DL, Shin SS, Bangma CH, Wei JT, et al. The prostate health index selectively identifies clinically significant prostate cancer. J Urol 2015;193:1163-9.
9. Roobol MJ, Schröder FH, van Leeuwen P, Wolters T, van den Bergh RC, van Leenders GJ, et al. Performance of the prostate cancer antigen 3 (PCA3) gene and prostate-specific antigen in prescreened men: exploring the value of PCA3 for a first-line diagnostic test. Eur Urol 2010;58:475-81.
10. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-9.
11. Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 2019;292:331-42.
12. Nir G, Karimi D, Goldenberg SL, Fazli L, Skinnider BF, Tavassoli P, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. JAMA Netw Open 2019;2:e190442.
13. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Methodol 1958;20:215-32.
14. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
15. Breiman L. Random forests. Mach Learn 2001;45:5-32.
16. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco (CA), USA. New York: ACM; 2016:785-94.
17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Neural Information Processing Systems Foundation, Inc.; 2017:3146-54.

18. Akaza H, Onozawa M, Hinotsu S. Prostate cancer trends in Asia. World J Urol 2017;35:859-65.

19. Droz JP, Albrand G, Gillessen S, Hughes S, Mottet N, Oudard S, et al. Management of prostate cancer in elderly patients: recommendations of a Task Force of the International Society of Geriatric Oncology. Eur Urol 2017;72:521-31.

20. Takeuchi T, Hattori-Kato M, Okuno Y, Iwai S, Mikami K. Prediction of prostate cancer by deep learning with multilayer artificial neural network. Can Urol Assoc J 2019;13:E145-50.

21. Morash C. PSA density: the comeback kid? Can Urol Assoc J 2012;6:51-2.

22. Verma A, St Onge J, Dhillon K, Chorneyko A. PSA density improves prediction of prostate cancer. Can J Urol 2014;21: 7312-21.

23. Sfoungaristos S, Perimenis P. PSA density is superior than PSA and Gleason score for adverse pathologic features prediction in patients with clinically localized prostate cancer. Can Urol Assoc J 2012;6:46-50.

24. Hamilton AS, Albertsen PC, Johnson TK, Hoffman R, Morrell D, Deapen D, et al. Trends in the treatment of localized prostate cancer using supplemented cancer registry data. BJU Int 2011;107:576-84.

25. Liu D, Lehmann HP, Frick KD, Carter HB. Active surveillance versus surgery for low risk prostate cancer: a clinical decision analysis. J Urol 2012;187:1241-6.