Discrimination of Patients with Prostate Cancer from Healthy Persons Using a Set of Single Nucleotide Polymorphisms

Mir Davood Omrani[1], Hossein Mohammad-Rahimi[2], Abbas Basiri[1], Milad Fallahian[3], Rezvan Noroozi[4], Mohammad Taheri[5]*, Soudeh Ghafouri-Fard[6]**

Purpose: Prostate cancer is the second cancer diagnosed in males. It accounts for about 4% of cancer-related mortality in men. Several genetic polymorphisms in different genes have been identified that alter the risk of this kind of malignancy.

Materials and methods: We used the random forest (RF) algorithm for prediction of prostate cancer risk in Iranian population using 13 different single nucleotide polymorphisms (SNPs) in four genes (ANRIL, HOTAIR, IL-6 and IL-8). The samples were divided into a training set (n=320) and a test set (n=80) to evaluate the generalization power for training algorithm. For hyper-parameters tuning, we used randomized search with 5-fold cross-validation for the following hyper-parameters: (1) Number of trees or estimators in the forest (set from 3 to 500); (2) The maximum number of leaf nodes (set from 2 to 32); (3) The maximum number of features used for the best split (set from 5 to 13); and (4) Using bootstrap samples in the trees building (True or False). Accuracy, sensitivity, specificity, and F1-score in both training and test sets were reported.

Results: The most important SNP was ANRIL-rs1333048: A/A (Gini index= 0.096) followed by ANRIL-rs10757278: G/G (Gini index= 0.059). Training Dataset Outcomes were as follow: Accuracy: 0.896, Sensitivity: 0.85, Specificity: 0.944 and F1 Score: 0.891. Test Dataset Outcomes were as follow: Accuracy: 0.787, Sensitivity: 0.775, Specificity: 0.800 and F1 Score: 0.784. The AUC Scores were 0.966 and 0.841 for training and test datasets, respectively.

Conclusion: The proposed panels of SNPs can predict risk of prostate cancer in Iranian population with appropriate accuracy.

Keywords: prostate cancer, single nucleotide polymorphism, IL-8, HOTAIR, ANRIL

INTRODUCTION

Prostate cancer ranks second among the diagnosed cancer in males. It accounts for about 4% of cancer-related mortality in men[(1)]. A comprehensive study in Iranian patients has shown that 97% of all cases have been adenocarcinoma. The other defined pathologies have been malignant carcinoma and transitional cell carcinoma[(2)]. At early phases of cancer development, prostate cancer usually does not have any symptoms and progresses in an indolent manner, needing minimal or even no therapeutic intervention. During its course, it can cause difficult urination, increased frequency or urgency in urination, nocturia and urinary retention and back pain in advanced stages, the latter being caused by metastasis[(3)]. Genome wide association studies (GWAS) conducted in different populations have identified tens of genetic polymorphisms that confer risk of this malignancy [(4-6)]. We have recently assessed the role of a number of single nucleotide polymorphisms (SNPs) in different genes in conferring risk of prostate cancer in Iranian population. These SNPs were located in ANRIL (rs1333045, rs4977574, rs1333048 and rs10757278) [(7)], HOTAIR (rs12826786, rs1899663 and rs4759314)[(8)], IL-6 (rs1800795 and rs2069845)[(9)] and IL-8 (rs4073, rs2227306 and rs1126647)[(10)]. In the current study, we applied the random forest (RF) algorithm for prediction of risk of prostate cancer based on the genotyping results of these 13 distinct SNPs. RF algorithm is an ensemble learning method for supervised classification introduced by Breiman[(11)]. This nonparametric tree-based approach combines the concepts of adaptive nearest neighbors with bagging[(12)]. RF method has the ability to assess correlation and interaction among var-

[1]Urology and Nephrology Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[2]Dental Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[3]Faculty of Civil Engineering, Amirkabir University of Technology, Tehran, Iran.
[4]Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland.
[5]Skull Base Research Center, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[6]Department of Medical Genetics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
*Correspondence: Urology and Nephrology Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. mohammad_823@yahoo.com.
** Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland. s.ghafourifard@sbmu.ac.ir.
Received June 2020 & Accepted May 2021

Table 1. The frequency and distribution of various polymorphisms.

		Control	Prostate cancer	BPH
ANRIL-rs1333045	C/T	129	65	57
	C/C	57	30	25
	T/T	14	5	18
ANRIL-rs1333048	A/A	110	22	27
	A/C	50	55	32
	C/C	40	23	41
ANRIL-rs4977574	G/G	82	62	65
	A/G	82	33	24
	A/A	36	5	11
ANRIL-rs10757278	A/G	91	58	65
	G/G	84	21	20
	A/A	25	21	15
HOTAIR-rs12826786	C/T	108	37	48
	C/C	60	25	28
	T/T	32	38	24
HOTAIR-rs4759314	A/A	121	61	54
	A/G	77	38	44
	G/G	2	1	2
RORA-rs11639084	C/C	126	59	73
	C/T	63	41	23
	T/T	11	0	4
RORA-rs4774388	T/T	105	64	49
	C/T	75	29	41
	C/C	19	7	10
IL-6-rs2069845	A/G	97	54	44
	A/A	82	33	47
	G/G	21	13	9
IL-6-rs56588968	C/G	87	41	40
	G/G	77	30	32
	C/C	36	29	28
IL-8-rs4073	A/T	96	53	27
	T/T	63	34	51
	A/A	41	13	22
IL-8-rs2227306	C/T	92	52	43
	C/C	76	37	35
	T/T	32	11	22
IL-8-rs1126647	A/T	89	39	37
	A/A	72	32	46
	T/T	39	29	17

iables. Notably, RF can facilitate selection and ranking of variables by calculating variable importance values. These features potentiate RF for evaluation of genomic data and bioinformatics investigation[13].

MATERIALS and METHODS

We used the RF algorithm for prediction of prostate cancer using 13 different SNPs. The samples were divided into a training set (n=320) and a test set (n=80) for the purpose of generalizing the outcome of the training algorithm. For hyper-parameters tuning, we used randomized search with 5-fold cross-validation for the following hyper-parameters: (1) Number of trees or estimators in the forest (set from 3 to 500); (2) The maximum number of leaf nodes (set from 2 to 32); (3) The maximum number of features used for the best split (set from 5 to 13); and (4) Using bootstrap samples in the trees building (True or False). A total of 1000 combination of these hyper-parameters were evaluated on the validation sets. After fixing the hyperparameters, we retrained the whole training set again.

Totally, 20 percent of the data samples were used as test set. The samples were chosen randomly. For evaluation of the hyperparameters, one fifth of the training set was used as validation set for the 5-fold-cross validation.

We did not set a limitation on the maximum depth of the trees. Therefore, the nodes were further expanded until all leaves became pure or until all leaves contained fewer samples than the “min samples split” amount. To avoid overfitting, we used the k-fold cross-validation technique.

In the current study, we used the Python programming language version 3.8.2. For applying the RF algorithm and hyper-parameters randomized search, we implemented Python Scikit-Learn 0.23.0 (https://scikit-learn.org/).

Accuracy, sensitivity, precision, and F1-score in both training and test sets were reported. Precision, sensitivity and F1-score were defined in equations 1-3 [14]:

$$\text{Equation 1. } Specificity = \frac{TN}{TN+FP}$$

$$\text{Equation 2: } Sensitivity = \frac{TP}{TP+FN}$$

$$\text{Equation 3. } F1 - Score = \frac{2 .Precision .Recall}{Precison+Recall}$$

Where, TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

Furthermore, we used the receiver operating characteristic (ROC) curve and area under curve (AUC) score to evaluate the performance of the model. We also presented the most important SNPs based on the impurity-based feature importance (also known as the Gini importance). The Gini index measures the importance of a feature by computing the level of the impurity of samples assigned to a node based on a split at its parent [15]. Gini index was calculated using Equation 4:

$$\text{Equation 4: } i(\tau) = 1 - p_1^2 - p_0^2$$

$$\text{Considering, } p_k = \frac{n_k}{n}$$

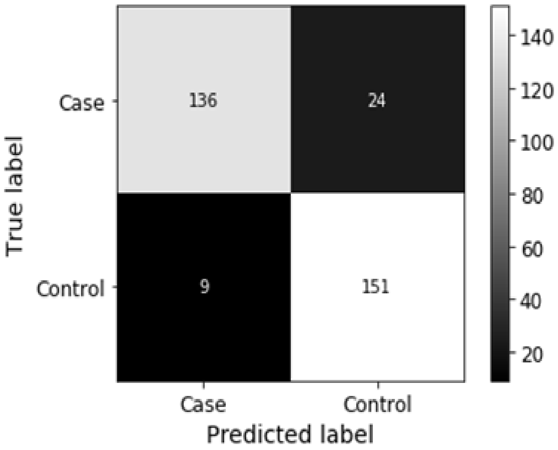

Figure 1. Training Dataset Confusion Matrix. The color bar next to the chart shows the frequency.

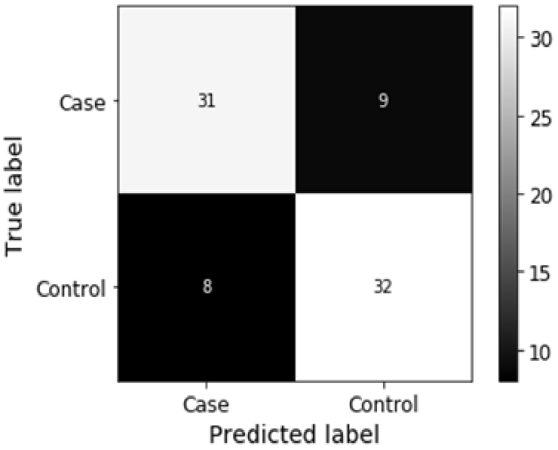

Figure 2. Test Dataset Confusion Matrix. The color bar next to the chart shows the frequency.

Where, *n* is the number of the total the total samples, nk is the number of samples from class *k = {0, 1}*, pk is the fraction of nk out of n samples at node τ.

We measured the generalization power based on the test data rather than the generalization. The strategy for setting the optimum value of hyperparameters (Hyperparameter Tuning) was randomized search and k-cross fold-validation.

RESULTS

Samples containing at least one NaN value were ruled out. The frequency and distribution of various polymorphisms are summarized in **Table 1**.

In the hyper-parameter tuning stage, hyper-parameters were set as follow: 1) Number of trees = 34; 2) The maximum leaf nodes = 30; 3) The maximum features = 8; and 4) Using bootstrap = True. The most important SNP was ANRIL-rs1333048: A/A (Gini index= 0.096)

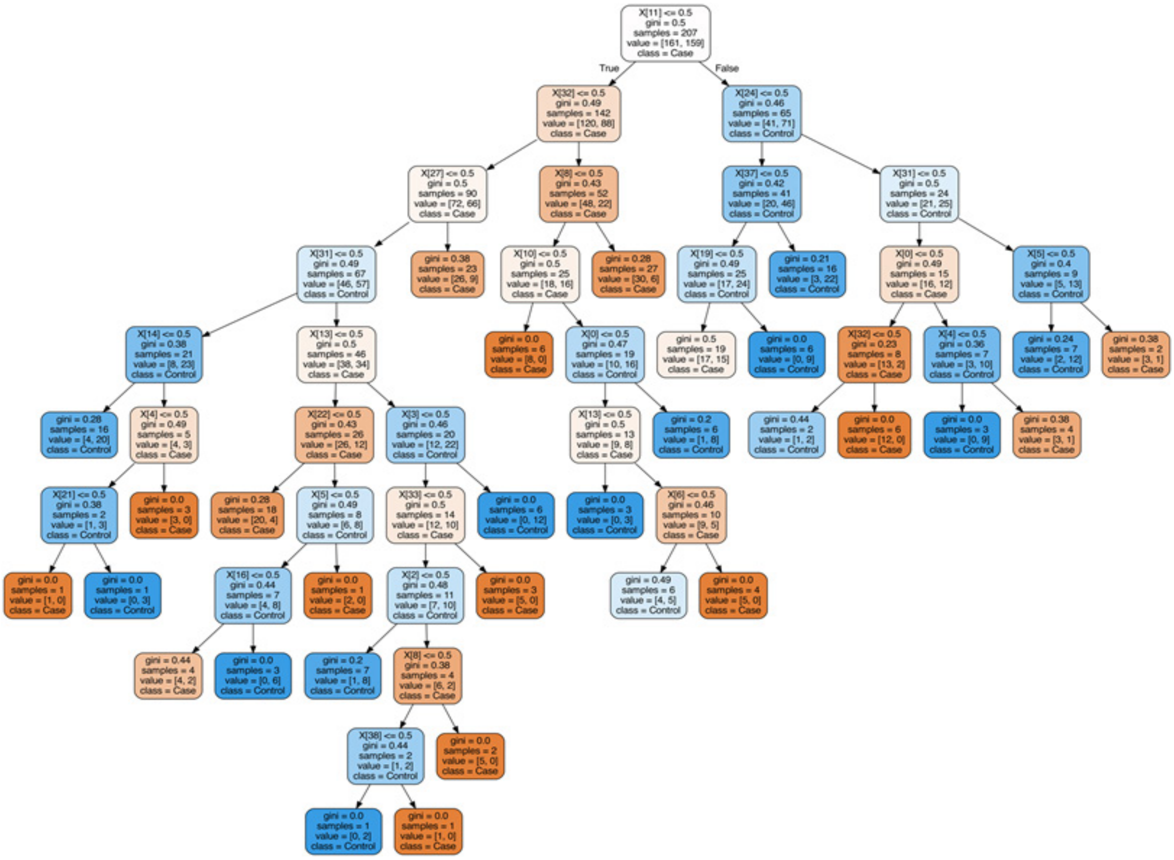
X[11] <= 0.5 gini = 0.5 samples = 207 value = [161, 159] class = Case
True
False
X[32] <= 0.5 gini = 0.49 samples = 142 value = [120, 88] class = Case
X[24] <= 0.5 gini = 0.46 samples = 65 value = [41, 71] class = Control
X[27] <= 0.5 gini = 0.5 samples = 90 value = [72, 66] class = Case
X[8] <= 0.5 gini = 0.43 samples = 52 value = [48, 22] class = Case
X[37] <= 0.5 gini = 0.42 samples = 41 value = [20, 46] class = Control
X[31] <= 0.5 gini = 0.5 samples = 24 value = [21, 25] class = Control
X[31] <= 0.5 gini = 0.49 samples = 67 value = [46, 57] class = Control
gini = 0.38 samples = 23 value = [26, 9] class = Case
X[10] <= 0.5 gini = 0.5 samples = 25 value = [18, 16] class = Case
gini = 0.28 samples = 27 value = [30, 6] class = Case
X[19] <= 0.5 gini = 0.49 samples = 25 value = [17, 24] class = Control
gini = 0.21 samples = 16 value = [3, 22] class = Control
X[0] <= 0.5 gini = 0.49 samples = 15 value = [16, 12] class = Case
X[5] <= 0.5 gini = 0.4 samples = 9 value = [5, 13] class = Control
X[14] <= 0.5 gini = 0.38 samples = 21 value = [8, 23] class = Control
X[13] <= 0.5 gini = 0.5 samples = 46 value = [38, 34] class = Case
gini = 0.0 samples = 6 value = [8, 0] class = Case
X[0] <= 0.5 gini = 0.47 samples = 19 value = [10, 16] class = Control
gini = 0.5 samples = 19 value = [17, 15] class = Case
gini = 0.0 samples = 6 value = [0, 9] class = Control
X[32] <= 0.5 gini = 0.23 samples = 8 value = [13, 2] class = Case
X[4] <= 0.5 gini = 0.36 samples = 7 value = [3, 10] class = Control
gini = 0.24 samples = 7 value = [2, 12] class = Control
gini = 0.38 samples = 2 value = [3, 1] class = Case
gini = 0.28 samples = 16 value = [4, 20] class = Control
X[4] <= 0.5 gini = 0.49 samples = 5 value = [4, 3] class = Case
X[22] <= 0.5 gini = 0.43 samples = 26 value = [26, 12] class = Case
X[3] <= 0.5 gini = 0.46 samples = 20 value = [12, 22] class = Control
X[13] <= 0.5 gini = 0.5 samples = 13 value = [9, 8] class = Case
gini = 0.2 samples = 6 value = [1, 8] class = Control
gini = 0.44 samples = 2 value = [1, 2] class = Control
gini = 0.0 samples = 6 value = [12, 0] class = Case
gini = 0.0 samples = 3 value = [0, 9] class = Control
gini = 0.38 samples = 4 value = [3, 1] class = Case
X[21] <= 0.5 gini = 0.38 samples = 2 value = [1, 3] class = Control
gini = 0.0 samples = 3 value = [3, 0] class = Case
gini = 0.28 samples = 18 value = [20, 4] class = Case
X[5] <= 0.5 gini = 0.49 samples = 8 value = [6, 8] class = Control
X[33] <= 0.5 gini = 0.5 samples = 14 value = [12, 10] class = Case
gini = 0.0 samples = 6 value = [0, 12] class = Control
gini = 0.0 samples = 3 value = [0, 3] class = Control
X[6] <= 0.5 gini = 0.46 samples = 10 value = [9, 5] class = Case
gini = 0.0 samples = 1 value = [1, 0] class = Case
gini = 0.0 samples = 1 value = [0, 3] class = Control
X[16] <= 0.5 gini = 0.44 samples = 7 value = [4, 8] class = Control
gini = 0.0 samples = 1 value = [2, 0] class = Case
X[2] <= 0.5 gini = 0.48 samples = 11 value = [7, 10] class = Control
gini = 0.0 samples = 3 value = [5, 0] class = Case
gini = 0.49 samples = 6 value = [4, 5] class = Control
gini = 0.0 samples = 4 value = [5, 0] class = Case
gini = 0.44 samples = 4 value = [4, 2] class = Case
gini = 0.0 samples = 3 value = [0, 6] class = Control
gini = 0.2 samples = 7 value = [1, 8] class = Control
X[8] <= 0.5 gini = 0.38 samples = 4 value = [6, 2] class = Case
X[38] <= 0.5 gini = 0.44 samples = 2 value = [1, 2] class = Control
gini = 0.0 samples = 2 value = [5, 0] class = Case
gini = 0.0 samples = 1 value = [0, 2] class = Control
gini = 0.0 samples = 1 value = [1, 0] class = Case

Figure 3. Visualization of the first estimator (Decision Tree) in our Random Forest Model

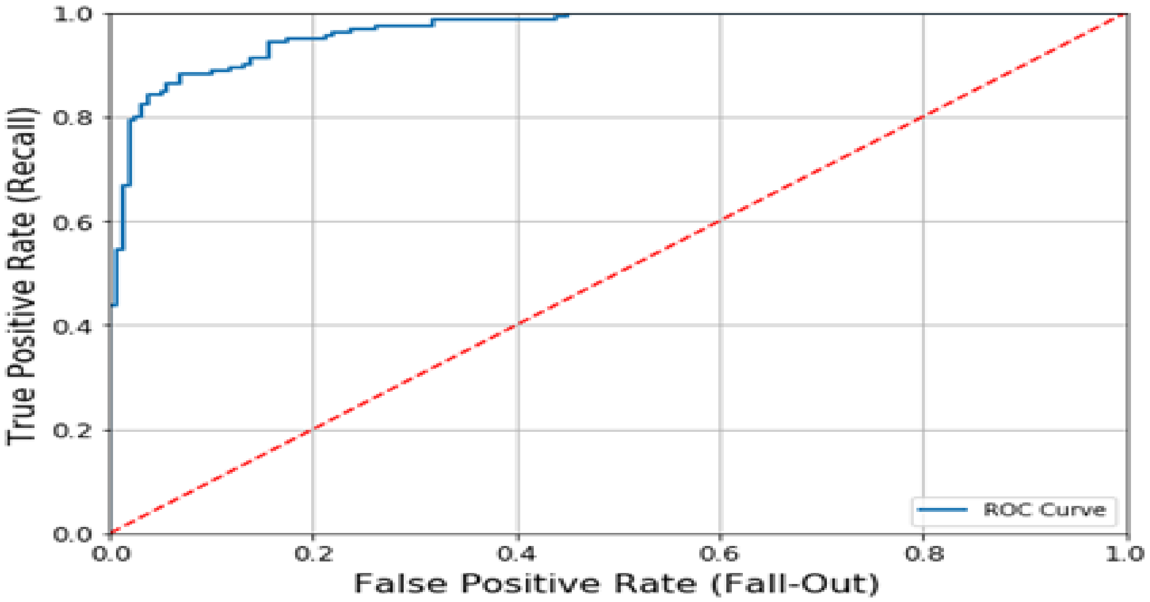

Figure 4. Training dataset ROC curve showing the AUC value of 0.966 for the proposed approach in the diagnosis of prostate cancer.

followed by ANRIL-rs10757278: G/G (Gini index= 0.059).

Training Dataset Outcomes were as follow: Accuracy: 0.896, Sensitivity: 0.85, Specificity: 0.944 and F1 Score: 0.891. Test Dataset Outcomes were as follow: Accuracy: 0.787, sensitivity: 0.775, Specificity: 0.800 and F1 Score: 0.784.

Figure 1 shows the Training Dataset Confusion Matrix. We also depicted Dataset ROC Curve for both training and test datasets (**Figures 4 and 5**). The AUC Scores were 0.966 and 0.841 for training and test datasets, respectively.

Features Importance of the assessed SNPs is shown in Figure 6. The best features have been demonstrated for ANRIL-rs1333048: A/A and ANRIL-rs10757278: G/G, respectively.

DISCUSSION

In the current study, we re-analyzed our genotyping data of 13 SNPs in a population of Iranian patients with prostate cancer using the RF method. This method has been previously applied in the analysis of SNPs in genetic studies. In GWAS, RF has been shown to be able in screening of SNPs with interaction effects. Such

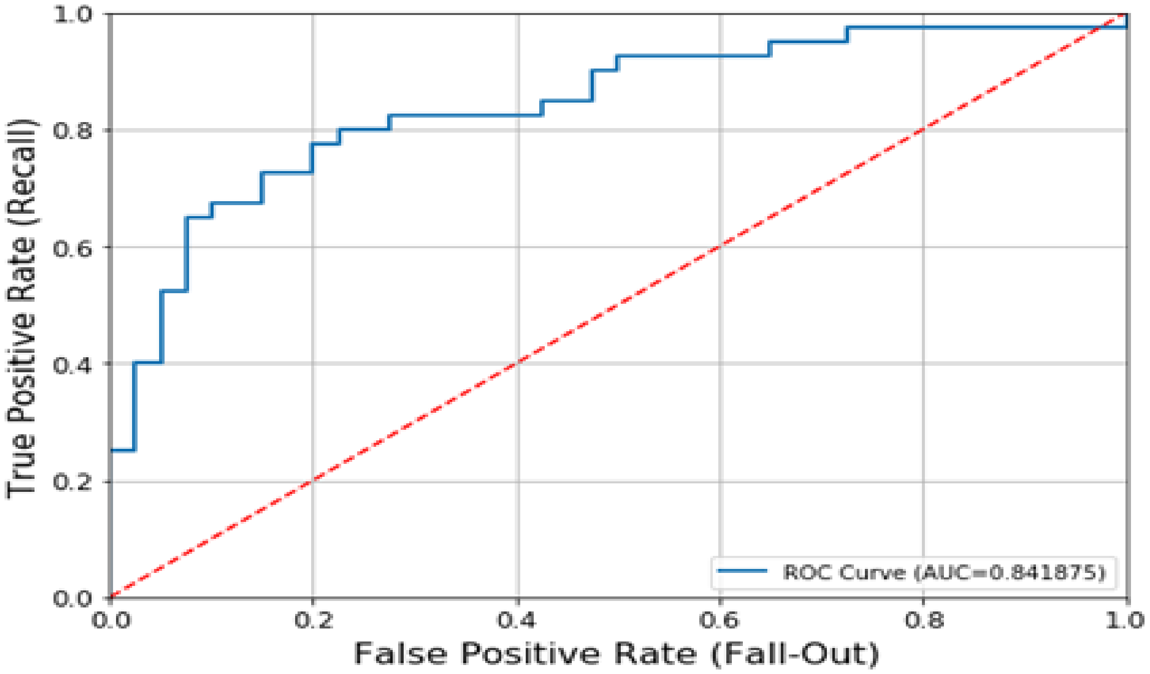

Figure 5. Test dataset ROC curve showing the AUC value of 0.841 for the proposed approach in the diagnosis of prostate cancer.

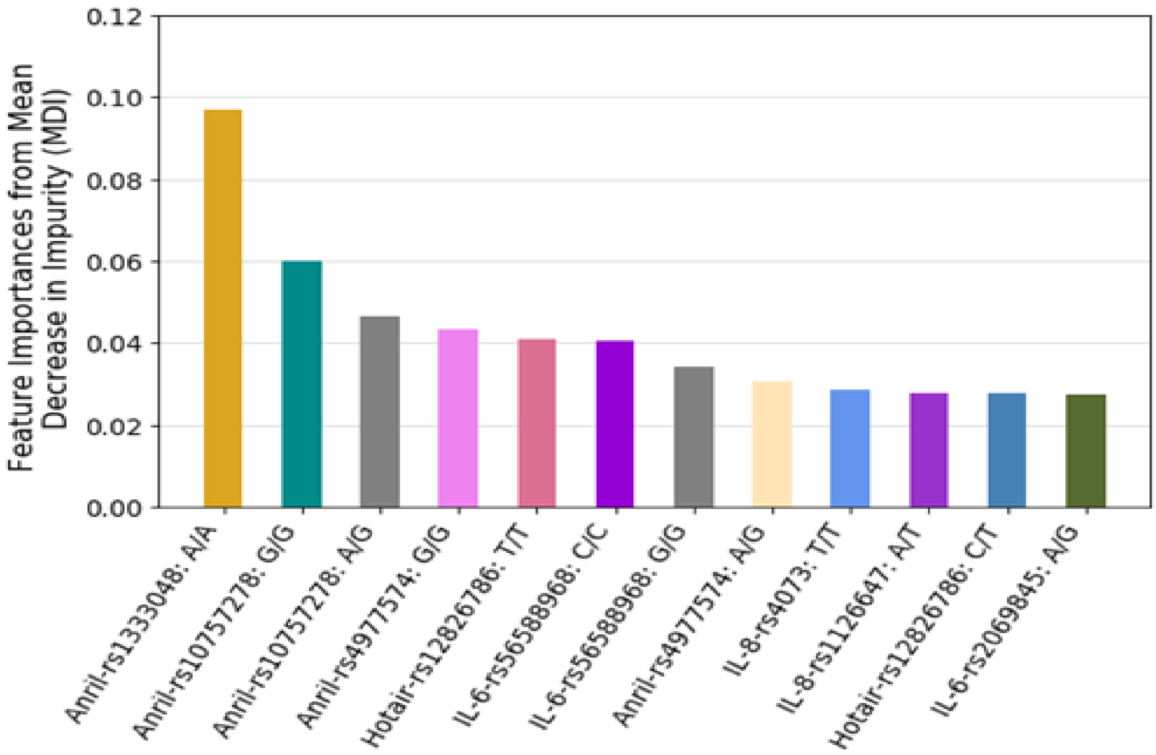

Figure 6. Features Importance of the assessed SNPs.

method has decreased the number of SNPs that should be recalled for additional study compared to routine univariate screening strategies[16]. RF has been successfully applied for assessment of the effects of 42 SNPs located in the asthma risk gene ADAM33 to reach 44% misclassification rate[17]. In coronary artery calcification, RF has been applied for predication of the effects of 287 tagged SNPs and 17 risk elements[18].

RF is superior to artificial neural network as it can decrease the high variance from a flexible model such as a decision tree through integrating several trees into one collaborative model. RF provides a different interpretation of a decision tree yet with superior performance. RF classifiers produce a large number of decision trees, without trimming or pruning. For each variable, this approach generates a significance score, which quantifies the variable relative contribution to prediction[19]. RF classifiers has been successfully used in various biomedical studies[20-22]. In a study by Masetic Z. et al[23], it has been reported that RF classifiers had better classification performance compared to decision tree, k-Nearest Neighbor, support vector machine, and artificial neural networks in congestive heart failure detection. In another study by Zahangir Alam Md. et al[21], it has been suggested that other classifiers, unlike RF, do not perform equally well over all used medical datasets. Similar to our study, they used k-fold cross validation for the model evaluation. The k-fold cross validation is a tool for evaluating a predictive model that splits the initial dataset into training sets and a validation sets for training and evaluating the model. It can also be used for the purpose for tuning the hyperparameters[24].

RF classifiers can also been usedfor analysis of the SNPs[19]. Regarding SNPs, numerous studies used RF algorithm for analysis of the SNPs[19,25-27]. Using RF, Van Dyke A. L, et al.[25] suggested that IL1A SNP is an important risk factor in predicting risk for non-small cell lung cancer among women using SNPs data. Staiano, A. et al[26] used RF algorithm to find SNPs associated with high cardiovascular risk.

RF has a valuable characteristic that enables a prompt calculable internal measure of variable importance. This feature can be applied to rank variables particularly in assessment of high-throughput genomic data. Node impurity indices (including the Gini index) are frequently used to appraise the importance measures [13]. In the current study, we calculated the Gini index importance according to the node impurity degree for node splitting. This approach led to the identification of the ANRIL-rs1333048: A/A (Gini index= 0.0967) and ANRIL-rs10757278: G/G (Gini index= 0.0599) genotypes as the most important genotypes in conferring risk of prostate cancer. The ANRIL rs1333048 SNPs have been previously shown to be associated with both generalized and localized aggressive periodontitis. Moreover, it resides in a common risk locus for coronary artery disease and periodontitis[28]. The GG genotype of rs10757278 has been remarkably associated with carotid plaque in female subjects[29]. The G allele of this SNP interferes with the binding site for STAT1. This SNP also alters expression of ANRIL and its nearby genes [30,31] in a way that the GG genotype confers the most decreased expression levels[30]. This SNP also affects alternative splicing of ANRIL[32]. Future studies are needed to unravel the molecular mechanisms leading to the importance of ANRIL rs1333048 and rs10757278 SNPs in the susceptibility to prostate cancer in the Iranian population.

Based on outcomes of training and test datasets accuracy, sensitivity, specificity and F1 score values were slightly lower in the test dataset. Moreover, the AUC scores were decreased in test dataset, albeit it remained significant. Thus, the proposed panels of SNPs can predict the risk of prostate cancer in the Iranian population with appropriate accuracy. This panel might be used as a screening panel for identification of at risk individuals. Further assessment of accuracy of this panel in lager cohorts of patients from different stages of prostate cancer might reveal its significance in the determination of disease course or prognosis.

ACKNOWLEDGMENTS

This study was financially supported by Urology and Nephrology Research Center.

CONFLICT OF INTEREST

The authors declare they have no conflict 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: a cancer journal for clinicians. 2018;68:394-424.
2. Basiri A, Eshrati B, Zarehoroki A, Golshan S, Shakhssalim N, Khoshdel A, et al. Incidence, Gleason Score and Ethnicity Pattern of Prostate Cancer in the Multi-ethnicity Country of Iran During 2008-2010. Urol J. 2020 May 4;17:602-6.
3. Rawla P. Epidemiology of Prostate Cancer. World J Oncol. 2019;10:63-89.
4. Eeles RA, Al Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nature genetics. 2013;45:385-91.
5. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nature genetics. 2014;46:1103-9.
6. Kote-Jarai Z, Al Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, et al. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nature genetics. 2011;43:785-91.
7. Taheri M, Pouresmaeili F, Omrani MD, Habibi M, Sarrafzadeh S, Noroozi R, et al. Association of ANRIL gene polymorphisms with prostate cancer and benign prostatic hyperplasia in an Iranian population. Biomarkers in medicine. 2017;11:413-22.
8. Taheri M, Habibi M, Noroozi R, Rakhshan A, Sarrafzadeh S, Sayad A, et al. HOTAIR genetic variants are associated with prostate cancer and benign prostate hyperplasia in an Iranian population. Gene. 2017;613:20-4.
9. Taheri M, Noroozi R, Rakhshan A, Ghanbari M, Omrani MD, Ghafouri-Fard S. IL-6 genomic variants and risk of prostate cancer. Urology journal. 2019;1:463-8.
10. Taheri M, Noroozi R, Dehghan A, Roozbahani GA, Omrani MD, Ghafouri-Fard S. Interleukin (IL)-8 polymorphisms and risk of prostate disorders. Gene. 2019;692:22-5.
11. Breiman L. Random Forests. Machine Learning. 2001 2001/10/01;45:5-32.
12. Breiman L. Bagging predictors. Machine learning. 1996;24:123-40.
13. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99:323-9.
14. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information processing & management. 2009;45:427-37.
15. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009 2009/07/10;10:213.
16. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC genetics. 2004;5:32.
17. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, et al. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2005;28:171-82.
18. Sun YV, Bielak LF, Peyser PA, Turner ST, Sheedy PF, Boerwinkle E, et al. Application of machine learning algorithms to predict coronary artery calcification with a sibship-based design. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008;32:350-60.
19. Meng Y, Yang Q, Cuenco KT, Cupples LA, DeStefano AL, Lunetta KL. Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks. BMC Proceedings. 2007 2007/12/18;1:S56.
20. Yang F, Wang H-z, Mi H, Cai W-w. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC bioinformatics. 2009;10:1-14.
21. Alam MZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked. 2019 2019;15:100180.
22. Mohapatra SK, Mohanty MN. Big data analysis and classification of biomedical signal using random forest algorithm. New Paradigm in Decision Science and Management: Springer; 2020. p. 217-24.
23. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Methods Programs Biomed. 2016 Jul;130:54-64.
24. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Statistics and

Computing. 2011;21:137-46.

25. Van Dyke AL, Cote ML, Wenzlaff AS, Chen W, Abrams J, Land S, et al. Cytokine and cytokine receptor single-nucleotide polymorphisms predict risk for non-small cell lung cancer among women. Cancer Epidemiol Biomarkers Prev. 2009 Jun;18:1829-40.

26. Staiano A, Di Taranto MD, Bloise E, D'Agostino MN, D'Angelo A, Marotta G, et al. Investigation of Single Nucleotide Polymorphisms Associated to Familial Combined Hyperlipidemia with Random Forests. In: Apolloni B, Bassis S, Esposito A, Morabito FC, editors. Neural Nets and Surroundings: 22nd Italian Workshop on Neural Nets, WIRN 2012, May 17-19, Vietri sul Mare, Salerno, Italy. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 169-78.

27. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005;21:2185-90.

28. Schaefer AS, Richter GM, Groessner-Schreiber B, Noack B, Nothnagel M, El Mokhtari N-E, et al. Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis. PLoS Genet. 2009;5:e1000378-e.

29. Zivotić I, Djurić T, Stanković A, Djordjević A, Končar I, Davidović L, et al. 9p21 locus rs10757278 is associated with advanced carotid atherosclerosis in a gender-specific manner. Exp Biol Med (Maywood). 2016;241:1210-6. PubMed PMID: 26941057. Epub 03/03. eng.

30. Liu Y, Sanoff HK, Cho H, Burd CE, Torrice C, Mohlke KL, et al. INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. PloS one. 2009;4.

31. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. PLoS Genet. 2010;6.

32. Zhang W, Chen Y, Liu P, Chen J, Song L, Tang Y, et al. Variants on chromosome 9p21. 3 correlated with ANRIL expression contribute to stroke risk and recurrence in a large prospective stroke population. Stroke. 2012;43:14-21.