JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 30, Number 12, 2023
© Mary Ann Liebert, Inc.
Pp. 1305–1314
DOI: 10.1089/cmb.2023.0040

A Highly Accurate Model for Screening Prostate Cancer Using Propensity Index Panel of Ten Genes

SHIPRA JAIN,* KAWAL PREET KAUR MALHOTRA,*
SUMEET PATIYAL,* and GAJENDRA PAL SINGH RAGHAVA

HIGHLIGHTS

- Application of machine learning techniques to identify biomarkers for PRAD cancer.
- Development of highly accurate models for classifying prostate cancer versus a normal sample.
- Introduction of the propensity index concept for enhancing model performance.
- Identification of top 10 genes using feature selection techniques.

ABSTRACT

Prostate-specific antigen (PSA) is a key biomarker commonly used to screen patients for prostate cancer. A significant number of unnecessary biopsies are performed every year due to poor accuracy of PSA-based biomarkers. In this study, we aim to identify alternate biomarkers based on gene expression that can be used to screen prostate cancer with higher accuracy. Our proposed machine learning model was trained and then tested on gene expression profiles of 500 prostate cancer and 51 normal samples in a 70:30 ratio. Numerous feature selection techniques have been used in this study to identify potential biomarkers. These identified genes have been used to develop various machine learning models for distinguishing between prostate cancer samples and healthy controls. Our logistic regression-based model achieved the highest area under the curve (AUC) of 0.91 with accuracy of 82.42% on the validation dataset. We introduced a new approach called propensity index, where expression of the gene is converted into propensity. Our propensity-based approach significantly improved the performance of classification models and achieved an AUC of 0.99 with accuracy of 96.36% on the validation dataset. We also identified and ranked biomarker genes that can be used to distinguish prostate cancer patients from healthy individuals with high accuracy. It was observed that single-gene-based biomarkers can only achieve accuracy of around 90%. In this study, we achieved the best performance using a panel of 10 genes; a random forest model using the propensity index was used. With rapid advancement, we hope that our proposed gene panel will be implemented for identifying and screening of prostate cancer, avoiding biopsy procedures.

Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, New Delhi, India.
*These authors contributed equally to this work.

Keywords: Gene Biomarker, machine learning techniques, PRAD cancer biomarker, propensity index, prostate cancer.

1. BACKGROUND

Prostate Adenocarcinoma (PRAD) is the second most prevalent cancer diagnosed in men around the world (Rawla, 2019). With recent advancements in health care, patients with prostate cancer are undergoing clinical biopsy procedures for proper diagnosis and management of the disease. A better understanding of the molecular mechanisms responsible for the onset of prostate carcinogenesis would help in exploring novel therapeutic methods.

In the literature, the prostate-specific antigen (PSA) test is a widely used test for detecting prostate cancer at a clinically significant stage for better treatment outcomes (Kirby, 2016). Higher PSA levels could indicate benign prostatic enlargement at an early stage. Due to false-positive prediction by this test, it leads to overdiagnosis and many unnecessary biopsies (Mengual et al., 2015). In one of the studies, serological FSCN1 levels were quantified to identify correlation with the disease, but were not found to be significant among both groups (Tătaru et al., 2021).

Recently, two urine-based RNA biomarkers, prostate cancer antigen 3 (PCA3) (Hessels et al., 2003) and fusion of two genes *TMPRSS2:ERG* (Laxman et al., 2006), have also been reported, which can be used to distinguish between men with early-stage disease and men in high-risk stages. Studies have reported the molecular mechanisms involved in development of PRAD, such as fusion of the members of the E26 transformation-specific family of transcription factors with androgen-regulated promoters (e.g., *TMPRSS2*) (Tomlins et al., 2009) and occurrence of point mutations of *TP53*, *FOXA1*, *PTEN*, and *SPOP* genes (Barbieri et al., 2012).

PCA3 (originally named as differential display code 3) is a urine-based biomarker that is widely used for prostate cancer detection (Bussemakers, 1999). Apart from genomic changes, epigenetic level changes have also been reported in cases of prostate cancer such as *GSTP1* hypermethylation in up to 70% of cases (Brooks et al., 1998; Ferro et al., 2017).

In one study, researchers claimed to identify a three-gene panel (*HOXC6*, *TDRD1*, and *DLX1*) as a promising tool to identify men with prostate cancer even though they have been reported to have low serum PSA (sPSA) values (Leyten et al., 2015). Researchers proposed a method, SelectMDx, which analyzes RNA-based biomarkers, *HOXC6* and *DLX1*, through reverse transcription to reduce the need for an initial biopsy test (Haese et al., 2019). This method is applied on post-digital rectal examination (DRE) patients and measures the *HOXC6* and *DLX1* mRNA levels (Carlsson and Roobol, 2017).

In one study, researchers proposed the ConfirmMDx method, which is a tissue-based epigenetic test developed in a study of 350 men with negative biopsy or repeat biopsy in the last two years (Partin et al., 2014). The test builds on a ''field effect'' phenomenon (Stewart et al., 2013). Due to limited data and availability of samples, this method is not regularly recommended in clinical practice.

In recent studies on better clinical management of cancer, the use of machine learning techniques has contributed to early detection of cancer (Camacho et al., 2018; Tătaru et al., 2021). There is need to identify reliable biomarkers for screening of prostate cancer to avoid unnecessary biopsies (Cucchiara et al., 2018). This motivated us to design this study for identifying biomarkers for screening prostate cancer patients with high precision.

In this study, we aimed to identify gene expression-based biomarkers that distinguish between prostate cancer patients and healthy controls. To select relevant features, we introduce single-gene-based feature selection techniques such as mean, significance difference in mean, and area under the curve (AUC). These techniques allow us to rank genes based on their classification efficiency. We selected the top 10 genes using each feature selection technique.

We implemented seven machine learning techniques to develop prediction models using selected gene features for identification of prostate cancer patients. To improve performance, we proposed a propensity index-based approach for developing prediction models using propensity instead of expression of genes. This propensity index method significantly improved the performance of models.

2. METHODS

2.1. Dataset

We downloaded the GDC TCGA prostate cancer (PRAD) dataset using the Xena Browser (https://xenabrowser.net/datapages/), which contains the gene expression profiles of 500 prostate cancer samples and 51 normal samples. It contains expression of 20,530 genes for each sample. In this study, fragments per kilobase of transcript per million mapped reads (FPKM) values of RNA transcripts are used as quantification values. Due to large variation in FPKM values, we normalized values using log2 after addition of 1.0 as a constant number to each of the FPKM values.

2.2. Feature selection techniques

In this study, we used three types of feature selection techniques based on mean, significance difference in mean, and AUC. We applied these approaches to identify genes whose expression could be used to easily distinguish between prostate and nonprostate subjects. We extracted the top 10 genes using these feature selection approaches from a list of 20,530 gene identifiers. A brief description of each technique is given.

2.2.1. Mean-based approach. In this approach, we calculated the mean expression of each gene for prostate cancer patients as well as for healthy samples. Then, we computed the difference in mean expression between prostate cancer and healthy samples for each gene. If the difference is high, it means that the gene can be used to distinguish between the two types of samples. We ranked genes based on the difference in mean and selected the top genes with the maximum difference.

The following formula has been used for computing the difference in mean for a given gene:

$$D_g = \left| \mathrm{Mean}\left(\mathrm{PC}_g\right) + \mathrm{Mean}\left(\mathrm{NPC}_g\right) \right| \tag{1}$$

where D_g represents the difference in mean for gene g, PC_g represents the expression of gene g in prostate cancer samples, and NPC_g represents the expression of gene g in nonprostate cancer samples.

Gene identifiers were sorted in decreasing order of the absolute difference in mean values. The top 10 gene identifiers with the highest mean difference between prostate cancer and nonprostate cancer samples were selected from the sorted list.

2.2.2. Significance difference in mean. In this approach, we compute the level of significance in mean expression of a gene in prostate and nonprostate cancer samples. In addition to mean, we also compute the standard deviation in expression of a gene in prostate and nonprostate cancer samples.

The following formula is used to compute significance difference in mean for a given gene:

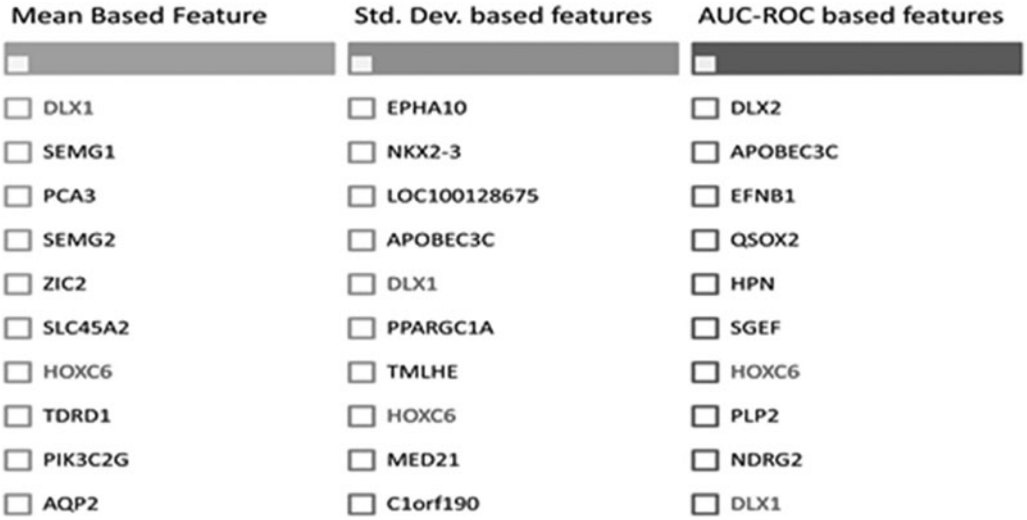
$$\mathrm{SD}_g = \frac{\left| \mathrm{Mean}\left(\mathrm{PC}_g\right) + \mathrm{Mean}\left(\mathrm{NPC}_g\right) \right|}{\mathrm{STD}\left(\mathrm{PC}_g\right) + \mathrm{STD}\left(\mathrm{NPC}_g\right)} \tag{2}$$

where SD_g represents the significance difference in mean expression of gene g in prostate and nonprostate cancer samples; PC_g represents the expression of gene g in prostate cancer samples; NPC_g represents the expression of gene g in nonprostate cancer samples; STD represents the standard deviation; $\mathrm{STD}(\mathrm{PC}_g)$ represents the standard deviation in expression of gene g in prostate cancer samples; and $\mathrm{STD}(\mathrm{NPC}_g)$ represents the standard deviation in expression of gene g in nonprostate cancer samples.

The genes are sorted in decreasing order of SD_g, that is, the value calculated by dividing mean by standard deviation. The top 10 gene identifiers with the highest difference between prostate and nonprostate cancer samples were selected from the sorted list.

2.2.3. Area under the curve. In this feature selection technique, we compute the discrimination power of each gene in terms of AUC (or AUC-receiver operating characteristic curve [ROC]). First of all, we calculated the mean expression of each gene ID for prostate cancer and noncancer data. The classification of samples is performed based on whether the expression of a given gene is above or below the threshold value.

Second, different threshold values are used to compute the AUC from the curve between the true-positive rate and false-positive rate. This process is performed for all genes in the dataset. Finally, the top 10 genes were selected, having maximum discrimination power in terms of AUC (Fig. 1).



Mean Based Feature	Std. Dev. based features	AUC-ROC based features
□ DLX1	□ EPHA10	□ DLX2
□ SEMG1	□ NKX2-3	□ APOBEC3C
□ PCA3	□ LOC100128675	□ EFNB1
□ SEMG2	□ APOBEC3C	□ QSOX2
□ ZIC2	□ DLX1	□ HPN
□ SLC45A2	□ PPARGC1A	□ SGEF
□ HOXC6	□ TMLHE	□ HOXC6
□ TDRD1	□ HOXC6	□ PLP2
□ PIK3C2G	□ MED21	□ NDRG2
□ AQP2	□ C1orf190	□ DLX1

FIG. 1. List of genes selected based on different feature selection techniques; top 10 genes from each technique.

2.2.4. Propensity index matrix. In this study, we coined the concept of the propensity index matrix for feature extraction from gene expression. In this method, we computed the range of gene expression for a given gene in our dataset and the difference was divided into 10 equal bins. For samples in each bin, we compute the propensity score for prostate and nonprostate cancer.

In the next step, we replaced the expression value of a gene by a propensity score based on the bin it belongs to. A new dataset is created using the propensity index score, which is provided as an input file in machine learning techniques for classification models.

2.3. Application of machine learning techniques

In this study, we applied seven different machine learning techniques for developing classification models. These techniques are support vector machine, K-nearest neighbor, decision tree, random forest (RF), linear regression, Gaussian Naive Bayes, and XGBoost machine learning. These techniques have been implemented using the Python library, scikit-learn.

2.4. Evaluation of models

In this study, we used cross-validation techniques to evaluate the performance of our models. We randomly divided our dataset into two datasets in the ratio of 70:30, where 70% of data are used for training and 30% are used for validation. We trained and tested our models on the training dataset using a fivefold cross-validation technique, where four folds are used as the training dataset and the remaining one fold as the testing dataset.

This process of dividing training and testing datasets is repeated five times. The performance evaluation of developed models on the testing dataset is called internal validation. To optimize the performance of our models on the training dataset, we optimized parameters. For the final optimized model, the best performance in internal validation was used to test an independent or validation dataset.

To measure the performance of our models, we used standard parameters commonly used to measure the performance of classification models. Both threshold-dependent and threshold-independent parameters are reported to evaluate the performance. We computed sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) as threshold-dependent parameters using the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100 \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100 \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \times 100 \tag{6}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{7}$$

where FP is false positive, FN is false negative, TP is true positive, and TN is true negative.

The area under the receiver operating characteristic curve, commonly known as AUC or ROC, is reported as a standard parameter for threshold-independent measures.

3. RESULTS

We developed classification models for classifying prostate and nonprostate cancer samples using seven machine learning techniques. First, we identified 10 genes (*DLX1*, *SEMG1*, *PCA3*, *SEMG2*, *ZIC2*, *SLC45A2*, *HOXC6*, *TDRD1*, *PIK3C2G*, and *AQP2*) using mean-based feature selection techniques. These selected genes were used to build machine learning techniques based on classification models. The performance of all models is evaluated on training and validation datasets.

As shown in Table 1, our logistic regression-based model achieved the maximum AUC of 0.91 on the training as well as validation dataset.

Second, we extracted the top 10 genes (*EPHA10*, *NKX2-3*, *LOC100128675*, *APOBEC3C*, *DLX1*, *PPARGC1A*, *TMLHE*, *HOXC6*, *MED21*, and *C1orf190*) using significance difference in mean-based feature selection techniques. These top 10 genes were used to build classification models using machine learning techniques. The performance of these models was evaluated on training and validation/testing datasets.

As shown in Table 2, our support vector machine (SVM)-based model achieved the highest AUC of 0.92 on the training dataset and AUC of 0.89 on the validation dataset.

Finally, we used the AUC-based approach for feature selection, where the top 10 genes were selected based on their performance. These genes (*DLX2*, *APOBEC3C*, *EFNB1*, *QSOX2*, *HPN*, *SGEF*, *HOXC6*, *PLP2*, *NDRG2*, and *DLX1*) showed the highest performance in terms of AUC when we used the threshold-based model for prediction. As shown in Table 3, our K-means nearest neighbor (KNN)-based model obtained the maximum AUC of 0.92 on the training dataset and AUC of 0.91 on the testing dataset.

3.1. Models based on the propensity index

In this study, we added a new concept for developing classification models. Instead of using expression of a gene as input, we used propensity of a gene as input. To convert expression of a gene to the propensity index of a gene, we divide the range of expression into 10 bins. In the next step, we compute the propensity index for each bin. Finally, expression of a gene is converted into a propensity score based on the expression in a given bin.

We developed classification models using the top 10 genes selected using mean-based feature selection techniques. As shown in Supplementary Table S1, we obtained the maximum AUC of 1.0 on the training dataset and 0.91 on the testing dataset using the RF model. The performance of our models improved significantly when we used the propensity index instead of expression (See Table 1 and Supplementary Table S1).

Table 1. The Performance of Machine Learning Technique-Based Models Developed Using Top 10 Genes Selected Using a Mean-Based Approach

	Training dataset					*Validation dataset*				
Model	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*
GNB	97.41	83.33	96.10	0.90	0.78	94.63	68.75	92.12	0.82	0.59
KNN	98.85	75.00	96.63	0.87	0.79	97.32	75.00	95.15	0.86	0.73
SVM	94.82	88.88	94.29	0.92	0.73	85.91	75.00	84.85	0.80	0.45
DT	98.56	83.33	96.36	0.90	0.78	97.99	62.50	94.54	0.80	0.66
RF	99.42	72.22	96.62	0.86	0.80	97.98	62.50	94.54	0.80	0.66
XGB	98.56	72.22	96.10	0.85	0.76	95.95	62.50	92.73	0.79	0.58
LR	93.96	88.89	93.50	0.91	0.70	83.22	75.00	82.42	0.91	0.70

AUC, area under the curve; DT, decision tree; GNB, Gaussian Naive Bayes; KNN, K-means nearest neighbor; LR, logistic regression; MCC, Matthews correlation coefficient; RF, random forest; Sens, sensitivity; Spec, specificity; SVM, support vector machine; XGB, XGBoost.

Table 2. The Performance of Machine Learning Technique-Based Models Developed Using Top 10 Genes Selected Using Significance Difference in Mean

	Training dataset					*Validation dataset*				
Model	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*
GNB	97.41	91.97	96.88	0.95	0.83	95.30	81.25	93.94	0.88	0.70
KNN	98.56	86.11	97.40	0.92	0.85	95.30	62.50	92.12	0.79	0.56
SVM	95.40	88.89	94.80	0.92	0.74	90.60	87.50	90.30	0.89	0.62
DT	97.70	77.77	96.09	0.88	0.75	97.32	56.25	93.33	0.76	0.58
RF	97.98	77.78	96.35	0.88	0.77	97.31	75.00	95.15	0.86	0.72
XGB	97.99	80.56	96.36	0.89	0.79	96.64	81.25	95.15	0.89	0.74
LR	95.40	97.22	95.57	0.96	0.80	91.95	87.50	91.15	0.88	0.65

Similarly, we developed models based on the propensity index of 10 genes selected using significance difference in mean-based feature selection. As shown in Supplementary Table S2, the logistic regression-based model achieved highest performance with an AUC of 1.00 on the training dataset and AUC of 0.97 on the validation dataset. In comparison with Table 2, the performance of prediction models reported in Supplementary Table S2 shows an improvement after converting expression values to propensity index values.

Finally, we developed models using the propensity index of top 10 genes obtained using the AUC-based feature selection approach. As shown in Supplementary Table S3, the RF model obtained best performance with an AUC of 1.00 on the training dataset and 0.99 on the validation dataset. It is clear from the above results that performance models developed using the propensity index (Supplementary Tables S1–S3) achieved better performance than models developed using gene expression (Tables 1–3).

3.2. Single gene-based classification

To understand the importance of individual genes in distinguishing prostate and nonprostate samples, we developed threshold-based models that can be used to identify prostate cancer samples based on expression of a single gene. Thus, after identifying the best genes using feature selection techniques, we ranked them on the basis of their capability of accurately predicting prostate cancer. All 30 genes extracted using feature selection techniques (i.e., top 10 using the mean-based method, 10 using the standard deviation-based method, and 10 using the AUC-based method) are considered for ranking.

After removing duplicate genes, we ranked these genes based on probability of correct prediction of prostate cancer. As shown in Supplementary Table S4, 13 genes have prognostic capability with probability of correct prediction from 0.97 to 0.989. In Supplementary Table S4, we also added the performance of *KLK3*, a gene associated with PSA (a commonly used test). It is clear that the performance of the *KLK3* gene is very poor in comparison with other genes used in our study.

In addition to ranking of genes, we also determined whether the expression of these genes in prostate cancer samples is statistically significant or not. We plotted gene expression values of the genes as box plot figures using the Gene Expression Profiling Interactive Analysis tool (Tang et al., 2017).

Table 3. The Performance of Machine Learning Technique-Based Models Developed Using Top 10 Genes Selected Using an Area Under the Curve-Based Approach

	Training dataset					*Validation dataset*				
Model	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*	*Sens*	*Spec*	*Accuracy*	*AUC*	*MCC*
GNB	94.54	94.45	94.53	0.94	0.75	93.29	87.50	92.73	0.90	0.68
KNN	98.27	86.11	97.14	0.92	0.83	95.30	87.50	94.55	0.91	0.74
SVM	90.51	94.45	90.90	0.92	0.65	88.59	87.50	88.48	0.88	0.58
DT	97.99	83.34	96.10	0.91	0.80	96.64	68.75	93.94	0.83	0.65
RF	98.85	72.22	97.40	0.86	0.77	97.31	81.25	95.76	0.89	0.76
XGB	98.27	75.00	96.09	0.87	0.76	97.31	81.25	95.76	0.89	0.76
LR	92.53	94.45	92.72	0.93	0.70	88.59	87.50	88.48	0.88	0.58

As shown in Figure 2, the box plots generated depict that 7 of 14 genes (*HPN*, *HOXC6*, *DLX1*, *SGEF*, *EPHA10*, *TDRD1*, and *PCA3*) are significant.

4. DISCUSSION

In this study, we aimed to identify gene expression-based biomarkers that distinguish between prostate cancer patients and healthy controls. We proposed a machine learning-based model, in which features were extracted using mean-, significance difference in mean-, and AUC-based methods. We identified the top 10 genes based on these three methods and further applied machine learning techniques for screening of prostate cancer and healthy control subjects with high accuracy.

In this study, we converted expression values of top 10 genes identified by feature selection techniques into the propensity index matrix. Classification models were developed using the propensity matrix, which showed significant improvement over models developed using gene expression. This is a novel approach used in this study to improve the accuracy of prediction.

With the implementation of our method, we aim to prevent misinterpretation of clinical diagnostic tests, which may lead to unnecessary treatments such as surgery and radiation therapy. These treatments might expose subjects to unfavorable side effects such as incontinence, erectile dysfunction, infection, and pain.

Since PSA is the widely accepted primary blood test for prostate cancer detection, we explored the possibility of using the *KLK3* gene (PSA-associated gene) as a potential biomarker. Due to the small difference between mean expression values for normal and prostate cancer patients, the *KLK3* gene was not identified in the top 10 genes through the feature selection techniques used in the current study.

From Figures 2 and 3, it is also evident that the *KLK3* gene's performance as a biomarker for prostate cancer is not good in terms of sensitivity, specificity, and accuracy. Various feature selection techniques were applied to determine genes that can strongly distinguish between tumorous and nontumorous records. It was observed that *DLX1* and *HOXC6* appeared in the top 10 genes with all three feature extraction techniques.

In a recent study, researchers reported the SelectMDx method, which proposed *HOXC6* and *DLX1* as RNA-based urine biomarkers (Haese et al., 2019). They reported an AUC of 0.85 with 93% sensitivity, 47% specificity, and 95% negative predictive value and PCPTRC AUC of 0.76 on the validation cohort.

Another study reported a urinary three-gene panel, that is, *HOXC6*, *TDRD1*, and *DLX1*, as a tool to distinguish prostate cancer patients with low sPSA values (Leyten et al., 2015). They reported an AUC value of 0.77 for these three biomarkers.

We also developed a KNN model using these three biomarkers, ran the process 10 times by shuffling data each time, and achieved average AUCs of 0.927 ± 0.009 (mean $\pm$ SD) for training and 0.971 ± 0.002 for testing datasets. We also converted expression values to propensity index scores for the support vector classifier (SVC) model and obtained AUCs of 0.981 ± 0.002 for the training dataset and 0.914 ± 0.001 for the testing dataset. These results were highly unbalanced in terms of sensitivity and specificity.

In a recent study, authors reported that *RTN1*, *HLA-DMB*, and *MRI1* differentially expressed genes can classify prostate cancer samples (450 samples) into three laterality classes (left, right, and bilateral) with almost 99% accuracy (Hamzeh et al., 2020). In this study, members tried to predict the location of prostate tumor tissue in samples using machine learning methods.

Another study in the literature aims to predict potential genetic biomarkers for each Gleason score group. In this study, researchers considered the RNA-Seq dataset of 104 prostate cancer patients and reported PIAS3 as a potential biomarker for Gleason score $4+3=7$ and *UBE2V2* for Gleason score 6 (Hamzeh et al., 2019). They further validated their results using a dataset of 499 prostate cancer patients and reported overall accuracy of 93.3% for training data and 87% for the validation dataset.

In this study, we plotted box plots to understand the potential of the top 14 genes ranked based on probability of correct prediction. We reported that *HPN*, *HOXC6*, *DLX1*, *SGEF*, *EPHA10*, *TDRD1*, and *PCA3* are found to be significant. Using past studies, we mapped the role of *DLX1*, *TDRD1*, and *HOXC6* in prostate cancer. In the literature, we found studies that explain the role of *HPN* (Ma et al., 2020), *SGEF* (Wang et al., 2012), *EPHA10* (Nagano et al., 2014), and *PCA3* (Bussemakers, 1994) in prostate cancer diagnosis and prognosis.

These studies further validate our findings. In the current study, we applied three feature selection techniques, that is, mean-based, standard deviation-based, and AUC ROC-based approaches, for identifying the

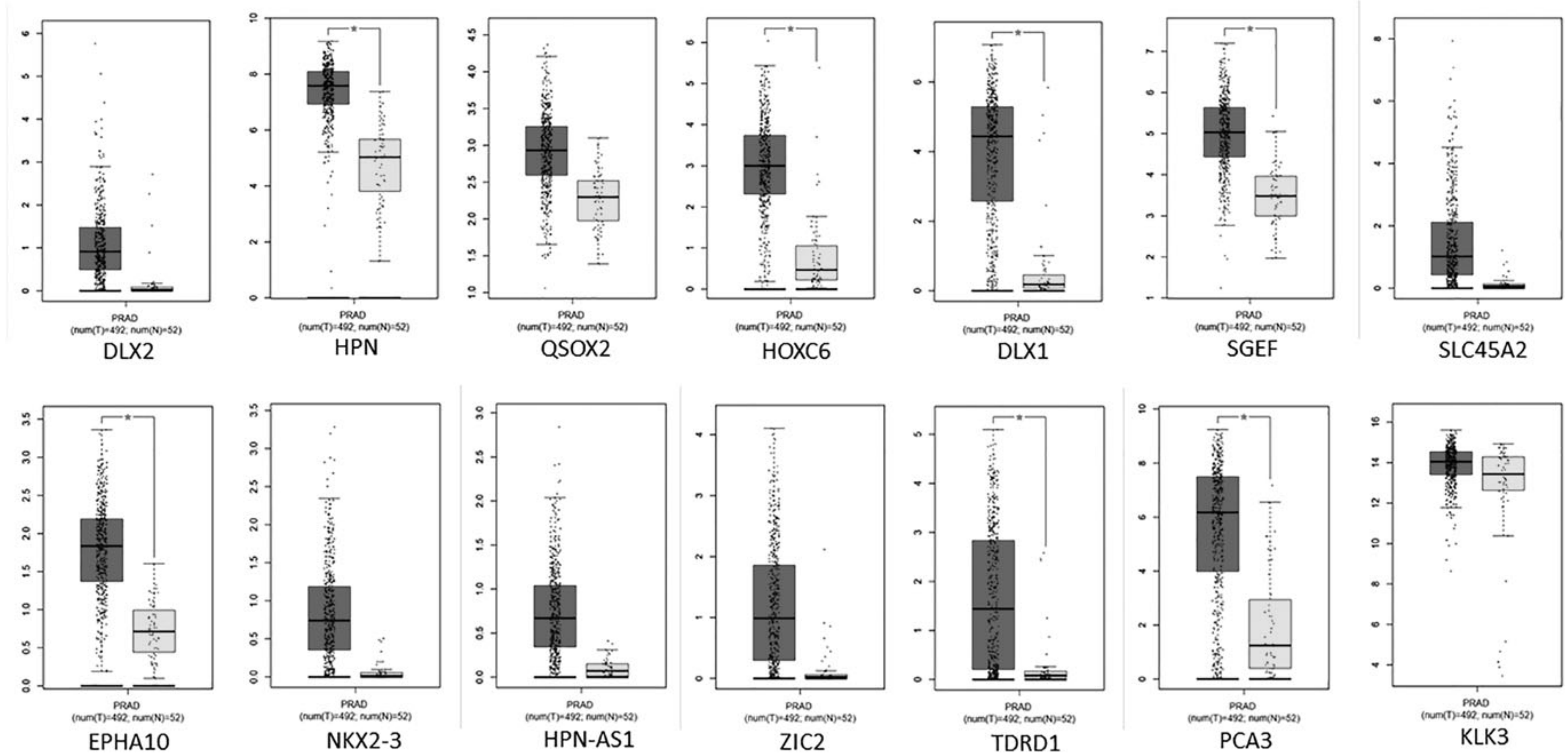


FIG. 2. Box plots of ranked genes. In this figure, the *dark* box plot depicts cancer samples and *light* depicts normal samples.

top 10 gene identifiers for classifying prostate cancer samples versus normal samples. Of the top 10 genes identified, *DLX1* and *HOXC6* were found to be present using all three approaches.

Furthermore, to understand the role of the *DLX1* and *HOXC6* genes in functional pathways, we explored Gene Ontology (GO) annotations of both genes. The *DLX1* gene, also known as distal-less homeobox 1, is a protein-encoding gene and is located on the long arm of chromosome 2. GO annotation of the *DLX1* gene includes sequence-specific DNA binding and chromatin binding. In the literature, DLX1 is reported to be associated with dental fluorosis and Witkop syndrome. It is involved in related pathways such as DNA damage/telomere stress-induced senescence and regulation of the nuclear *SMAD2/3* signaling pathway.

The *HOXC6* gene, also referred to as homeobox C6, is also a protein-coding gene and is located in a cluster on chromosome 12. The homeobox gene family usually encodes a highly conserved family of transcription factors involved in a crucial role (such as morphogenesis) in multicellular organisms. Furthermore, GO annotations include DNA-binding transcription factor activity and transcription corepressor activity.

Diseases that are reported to be linked with *HOXC6* include lymphoma, non-Hodgkin lymphoma, and familial lymphoma. With recent advancements, there is always scope for improvement. Apart from these approaches, various other measures such as entropy changes can be used to select genes that will lead to higher information gain. We could also apply network analysis models to establish connections between various gene IDs.

Building networks for tumorous and nontumorous gene expression data could provide deeper insights into the molecular mechanisms involved in development of cancerous conditions.

AUTHORS' CONTRIBUTIONS

S.J., K.P.K.M., and G.P.S.R. were involved in conception, design, and development of methodology. S.J. and K.P.K.M. were involved in acquisition of data. S.J., K.P.K.M., S.P., and G.P.S.R. were involved in analysis and interpretation of data and results. S.J., S.P., and G.P.S.R. were involved in writing, reviewing, and revision of the manuscript.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

The authors are thankful to the Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi (IIIT-Delhi) (DBT/2017/IIIT-D/834). S.P. is grateful to the Department of Biotechnology for providing fellowships.

SUPPLEMENTARY MATERIAL

Supplementary Table S1
Supplementary Table S2
Supplementary Table S3
Supplementary Table S4

REFERENCES

Barbieri CE, Baca SC, Lawrence MS, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 2012;44(6):685–689; doi: 10.1038/ng.2279

Brooks JD, Weinstein M, Lin X, et al. CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia. Cancer Epidemiol Biomarkers Prev 1998;7(6):531–536.

Bussemakers MJ, van Bokhoven A, Verhaegh GW, et al. DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res 1999;59(23):5975–5979; https://pubmed.ncbi.nlm.nih.gov/10606244/

Camacho DM, Collins KM, Powers RK, et al. Next-generation machine learning for biological networks. Cell 2018;173(7):1581–1592; doi: 10.1016/j.cell.2018.05.015

Carlsson S V. and Roobol MJ. Improving the evaluation and diagnosis of clinically significant prostate cancer in 2017. Curr Opin Urol 2017;27(3):198–204; doi: 10.1097/MOU.0000000000000382

Cucchiara V, Cooperberg MR, Dall'Era M, et al. Genomic markers in prostate cancer decision making. Eur Urol 2018;73(4):572–582; doi: 10.1016/j.eururo.2017.10.036

Ferro M, Ungaro P, Cimmino A, et al. Epigenetic signature: A new player as predictor of clinically significant prostate cancer (PCa) in patients on active surveillance (AS). Int J Mol Sci 2017;18(6); doi: 10.3390/IJMS18061146

Haese A, Trooskens G, Steyaert S, et al. Multicenter optimization and validation of a 2-gene MRNA urine test for detection of clinically significant prostate cancer before initial prostate biopsy. J Urol 2019;202(2):256–262; doi: 10.1097/JU.0000000000000293

Hamzeh O, Alkhateeb A, Zheng J, et al. Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. BMC Bioinformatics 2020;21(Suppl 2); doi: 10.1186/s12859-020-3345-9

Hamzeh O, Alkhateeb A, Zheng JZ, et al. A hierarchical machine learning model to discover gleason grade-specific biomarkers in prostate cancer. Diagnostics 2019;9(4); doi: 10.3390/diagnostics9040219

Hessels D, Klein Gunnewiek JMT, Van Oort I, et al. DD3PCA3-based molecular urine analysis for the diagnosis of prostate cancer. Eur Urol 2003;44(1):8–16; doi: 10.1016/S0302-2838(03)00201-X

Kirby R. The role of PSA in detection and management of prostate cancer. Practitioner 2016;260(1792):17–21, 3. https://pubmed.ncbi.nlm.nih.gov/27337755/

Laxman B, Tomlins SA, Mehra R, et al. Noninvasive detection of TMPRSS2:ERG fusion transcripts in the urine of men with prostate cancer. Neoplasia 2006;8(10):885–888; doi: 10.1593/neo.06625

Leyten GHJM, Hessels D, Smit FP, et al. Identification of a candidate gene panel for the early diagnosis of prostate cancer. Clin Cancer Res 2015;21(13):3061–3070; doi: 10.1158/1078-0432.CCR-14-3334

Ma X, Guo J, Liu K, et al. Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell RNA sequencing. Mol Cancer 2020;19(1); doi: 10.1186/s12943-020-01264-9

Mengual L, Musquera M, Ciudin A, et al. [Non-PSA serum markers for the diagnosis of PCa]. Arch Esp Urol 2015;68(3):229–239. https://pubmed.ncbi.nlm.nih.gov/25948796/

Nagano K, Yamashita T, Inoue M, et al. Eph receptor A10 has a potential as a target for a prostate cancer therapy. Biochem Biophys Res Commun 2014;450(1):545–549; doi: 10.1016/j.bbrc.2014.06.007

Partin AW, Van Neste L, Klein EA, et al. Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. J Urol 2014;192(4):1081–1087; doi: 10.1016/j.juro.2014.04.013

Rawla P. Epidemiology of prostate cancer. World J Oncol 2019;10(2):63–89; doi: 10.14740/wjon1191

Stewart GD, Van Neste L, Delvenne P, et al. Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: Results of the MATLOC study. J Urol 2013;189(3):1110–1116; doi: 10.1016/j.juro.2012.08.219

Tang Z, Li C, Kang B, et al. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res 2017;45(W1):W98–W102; doi: 10.1093/nar/gkx247

Tătaru OS, Martha O, Crocetto F, et al. Fascin-1 and its role as a serological marker in prostate cancer: A prospective case-control study. Future Sci OA 2021;7(9); doi: 10.2144/FSOA-2021-0051

Tătaru OS, Vartolomei MD, Rassweiler JJ, et al. Artificial intelligence and machine learning in prostate cancer patient management-current trends and future perspectives. Diagnostics (Basel) 2021;11(2):354; doi: 10.3390/DIAGNOSTICS11020354

Tomlins SA, Bjartell A, Chinnaiyan AM, et al. ETS gene fusions in prostate cancer: From discovery to daily clinical practice. Eur Urol 2009;56(2):275–286; doi: 10.1016/j.eururo.2009.04.036

Wang H, Wu R, Yu L, et al. SGEF is overexpressed in prostate cancer and contributes to prostate cancer progression. Oncol Rep 2012;28(4):1468–1474; doi: 10.3892/or.2012.1917

Address correspondence to:
Dr. Gajendra Pal Singh Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
A-302 R&D Block, Okhla Industrial Estate,
Phase III (Near Govind Puri Metro Station)
New Delhi 110020
India

E-mail: raghava@iiitd.ac.in