CANCER DIAGNOSIS & PROGNOSIS *5*: 720-734 (2025)
doi: 10.21873/cdp.10487

Robust Logistic Regression-based Diagnosis Method of Prostate Cancer Using Optimized Feature Selection on Race Specific Gene-expression Datasets

DAVID AGUSTRIAWAN[1], VINCENT KURNIAWAN[1], MARLINDA VASTY OVERBEEK[1], MOELJONO WIDJAJA[1], ADITHAMA MULIA[1], JHENO SYECHLO[1], MUHAMMAD IMRAN AHMAD[2], BESUT DARYANTO[3,4], KURNIA PENTA SEPUTRA[3,4], HERY SUSILO[3,4], EDVIN PRAWIRA NEGARA[3,4], REZA AKBAR EFFENDI[3,4] and SRINIVASULU YERUKALA SATHIPATI[5]

[1]Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia;
[2]Faculty of Intelligent Computing, Malaysia Perlis University, Arau, Malaysia;
[3]Department of Urology, Faculty of Medicine, Universitas Brawijaya, Malang, Indonesia;
[4]Saiful Anwar Hospital, Malang, Indonesia;
[5]Marshfield Clinic Research Institute, Marshfield Clinic Research Institute, Marshfield, WI, U.S.A.

Abstract

Background/Aim: Prostate cancer (PCa) incidence varies significantly by race, with Black men experiencing nearly 1.8 times higher prevalence than White men in the USA. Current prostate specific antigen (PSA)-based diagnostics lack specificity, and many machine learning models fail to consider racial differences in gene expression. This study proposes a race-aware PCa detection framework using optimized feature selection to improve diagnostic accuracy and fairness.
Materials and Methods: RNAseq-Count-STAR and clinical phenotype data from TCGA (554 patients) were analyzed. A feature selection pipeline integrating Differential Gene Expression analysis, Receiving Operating Characteristic (ROC) analysis, and Gene-Set Enrichment Analysis identified a 9-gene subset strongly associated with the PCa clinical pathway. The model was trained on White population data and validated on the Black population dataset using various data balancing techniques.
Results: The 9-gene logistic regression model achieved 95% accuracy in the White population and 96.8% accuracy in the Black population. Fairness analysis indicated minimal disparity between groups (4% difference in demographic parity, p=0.518). These results highlight the predictive value of race-specific biomarkers and demonstrate that biologically informed feature selection improves both accuracy and interpretability.
Conclusion: This study introduces a race-specific PCa detection framework that improves diagnostic accuracy using targeted biomarkers. It addresses misclassification risks in race-agnostic models and emphasizes the need for race-aware gene expression in ML diagnostics. Beyond detection, it enables personalized treatment, advancing precision medicine in PCa care.

Keywords: Prostate cancer, gene expression, feature selection, logistic regression, machine learning, bioinformatics.

David Agustriawan, Ph.D., Faculty of Engineering and Informatics/Universitas Multimedia Nusantara, Jalan Scientia Boulevard Gading, Curug Sangereng, Serpong, Kabupaten Tangerang, Banten 15810, Malaysia. Tel: +62 87781535936, e-mail: david.agustriawan@umn.ac.id

Received August 1, 2025 | Revised August 27, 2025 | Accepted September 2, 2025

Introduction

Prostate cancer (PCa) is one of the most prevalent cancers in men, with approximately 1.6 million cases and 366,000 deaths annually (1). It ranks second in male cancer incidence after lung cancer, contributing to 1,276,106 new cases and 358,989 deaths (3.8% of all cancer-related male deaths) in 2018 (2). During its early-stage PCa sometimes exhibits no symptoms, making prompt identification and therapy more difficult (2). PCa rates vary considerably according to race, ethnicity and geography, with black people in the US having a nearly 1.8-fold higher PCa rate than white people (3, 4). Existing diagnostic methods such as prostate specific antigen (PSA) testing face problems of specificity, leading to a high incidence of false positives and false negatives (5). Screening for PSA is associated with over- diagnosis and unnecessary treatment complications (6). For example, the use of a 3.0 ng/ml PSA threshold results in a 2.6% false negative rate for all PCa and 0.5% for clinically specific PCa (7). Given these constraints, alternative approaches to improving the accuracy of detection are being explored, with machine learning (ML) emerging as a promising solution (8). ML algorithms can improve predictive modelling by detecting complex patterns in data and outperform traditional PSA testing in terms of accuracy (9, 10). The integration of ML into diagnostic frameworks aims to reduce false positives, improve screening effectiveness and enable early intervention (11). This shift from traditional to data-driven diagnostics represents a significant advancement in PCa detection, potentially improving survival rates and patient outcomes (12). The rise of genomic data further enhances PCa detection through gene expression profiling where variations in DNA sequences can reveal disease risks and inform treatment strategies (13). Leveraging genomic insights also enables precision medicine, offering personalized interventions beyond the capabilities of imaging or PSA testing. This approach not only enhances early detection but also optimizes treatment responses, improving patient quality of life and minimizing mistreatment (14, 15).

Diagnostic models for PCa have been developed using a variety of methods, including advanced imaging, microarray analysis and clinical data collection ML techniques are increasingly used for their ability to analyze complex data sets and improve the accuracy of classification. Logistic regression, a widely used ML algorithm for binary classification, was used in several comparative PCa studies. In model (16) logistic regression reached an accuracy of 0.91, second only to multiple linear regression (0.96). The model (17) reported an area under the curve (AUC) score of 0.77, which was higher than the PSA score (0.67). Similarly, model (18) evaluated different ML techniques on clinical data and identified the multi-layer perception (MLP) as the best performer with an accuracy of 0.97. The gene expression-based models further improved the detection of PCa. Model (19) analyzed datasets containing genes, exon, exon cross-linking and isoform data and applied dimensionality reduction techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). The PCA used for the combined datasets resulted in the highest accuracy and precision. Model (20) used K-Nearest-Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Decision Tree Classifier (DTC) models in combination with a selection of variables using a signal-to-noise ratio (SNR) and a correlation coefficient. Their model achieved 100 percent accuracy in classifying the four selected genes and 85 percent accuracy in classifying PCa. Despite promising results, current models do not account for differences in gene expression between racial groups. The PCa biomarkers vary between populations due to genetic variation, which is a key consideration for diagnostic accuracy. For example, autopsy studies showed a higher incidence of high grade prostatic intraepithelial neoplasia in African American men compared to European American men (21). Despite progress in the detection of PCa through gene expression, no published study has adopted a race-specific approach using gene expression data. While the model (19) and model (20) addressed high dimensionality and class imbalance, they

Table I. *Summary of related studies and contributions of this work.*

Research	Their findings	Our contribution
Mohammed *et al.* (16)	Used logistic regression for binary classification in PCa diagnosis, achieving 0.91 accuracy, but did not explore higher-dimensional or race-specific models.	Proposed a race-specific ML model for PCa diagnosis, incorporating differential gene expression (DGE) and enrichment analysis for optimized feature selection.
Busetto *et al.* (17)	Applied logistic regression with an AUC of 0.77, outperforming PSA (0.67), but lacked genomic data integration.	Improved classification accuracy by integrating gene expression data with race-specific consideration and advanced feature selection.
Erdem *et al.* (18)	Compared ML algorithms on clinical data, identifying MLP as the best performer with 0.97 accuracy, but did not use gene expression or race-specific approaches.	Employed ML with gene expression profiles and DGE to enhance diagnostic accuracy across racially diverse populations.
Casey *et al.* (19)	Analyzed gene, exon, and isoform data using PCA and RFE for dimensionality reduction but did not include racial variations or use DESeq2 for DGE.	Applied DESeq2 for DGE-based feature selection and SMOTE for class balancing, addressing both high dimensionality and racial bias.
Bouazza *et al.* (20)	Used KNN, SVM, LDA, and DTC with SNR and correlation coefficients to select variables, achieving high accuracy, but without race-specific analysis or DGE methods.	Developed a comprehensive race-aware pipeline using DESeq2, gene enrichment analysis, and SMOTE to ensure equitable and accurate PCa diagnosis across populations.

PCa: Prostate cancer; ML: machine learning; DGE: differential gene expression; AUC: area under curve; PSA: prostate specific antigen; MLP: multi-layer perceptron; PCA: principal component analysis; RFE: recursive feature elimination; KNN: K-Nearest-Neighbors; SVM: support vector machine; LDA: linear discriminant analysis; DTC: decision tree classifiers.

did not include racial differences or use a differential gene expression analysis (DGE) approach using the DESeq2 tool to select the characteristics. The model (19) used SMOTE to balance the class, but did not consider race-related factors, which may limit clinical relevance. To bridge this gap, we propose a race-specific ML model for PCa diagnosis, which integrates the optimization of the selection of features by means of DGE and enrichment analysis of the gene pool. In addition, balancing techniques such as SMOTE will be used to improve the performance of the model and to ensure more accurate and equitable diagnostic results in a variety of populations. For more details, please check Table I.

This study proposes a robust logistic regression approach with optimized feature selection for PCa diagnosis with the goal of improving early detection and removing bias between race specific biomarkers, which further enhances its accuracy. This approach improves effectiveness and robustness by separating data by race, eradicating race bias and utilizing optimized feature selection through gene-set enrichment analysis which later is being used for logistic regression modelling. The combination of race-specific datasets, optimized feature selection and logistic regression modelling minimizes bias and offers a robust diagnostic method for PCa, thereby enhancing its ability for early detection.

Materials and Methods

The pipeline of our study is described in Figure 1. The methods consist of data collection, data preprocessing, feature selection, data up sampling, data splitting, hyperparameter tuning, training & testing.

Data collection. This study retrieved data from the open-source TCGA database, hosted by the University of California, Santa Cruz, on August 25, 2024. Two key datasets were used in this research: the RNAseq STAR - Counts dataset and the GDC TCGA Phenotypes dataset. The RNAseq dataset contains gene expression information obtained through sequencing. The count data has been pre-normalized using the log2(count+1) method, providing greater depth into transcriptional activity across different genes. This normalization allows for the identification of genes associated with specific conditions, such as cancer.

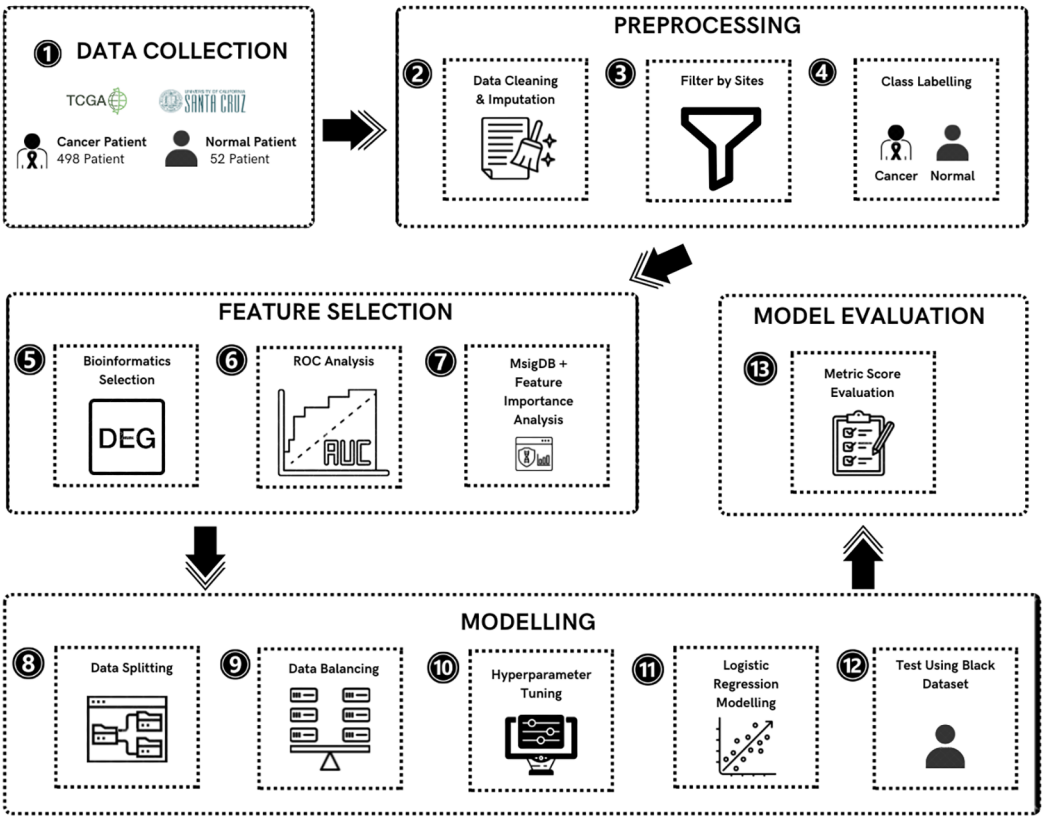

Figure 1. *Overview of the research pipeline utilized in this study.*

Preprocessing of gene expression data. First, both data need to be preprocessed separately first due to the different formats of data structure. For RNASeq STAR-Count we had to proceed with the data cleaning and imputation of missing values, and for the Phenotypes data we had to check thoroughly whether the feature is reliable and usable for the data merging later. After the data cleaning process, we matched the RNASeq STAR-Count data with phenotype data by matching its Ensembl_ID to filter the RNASeq STAR-Count to have the race specific dataset; for instance, we separated the data into White, Black, Native American, and Asian. We processed our data using Python version 3.12.7.

Feature selection. Features for model construction were selected using DGEs and Receiver Operating Characteristic (ROC) analysis to identify the most significant genes (22). DGE analysis was performed using PyDESeq2 version 0.4.10. The generated DGE list was filtered with cutoff thresholds of baseMean ≥10 and p-value <0.05 to exclude outlier genes, and those with baseMean <10 were excluded from further analysis. Genes were classified as up-regulated or down-regulated based on positive or negative $\log_2$ Fold Change, respectively (23). The filtered DGEs were then passed to ROC analysis to identify the most significant genes. A higher ROC value indicates a greater likelihood of true positive predictions, and genes with an area under the curve (AUC) >0.9 were selected, as previous studies suggest that genes with AUC values above this threshold are the most predictive for modeling (24). The selected genes were further refined through Gene-Set Enrichment Analysis (GSEA), with filtering against the

Table II. *Feature selection, data balancing and splitting scenario for this research.*

Feature selection	Train-test splitting	Balancing techniques
DGE (basemean ≥10 & *p*adj <0.05)	80/20 70/30 60/40	No Balancing, RandomUnderSampler, SMOTE, RandomOversampler, TOMEKLinks, BorderlineSMOTE, ADASYN, SVMSMOTE, KMeansSMOTE
DGE (basemean ≥10 & *p*adj <0.05 & log2FoldChange >0.35)	80/20 70/30 60/40	No Balancing, RandomUnderSampler, SMOTE, RandomOversampler, TOMEKLinks, BorderlineSMOTE, ADASYN, SVMSMOTE, KMeansSMOTE
DGE (basemean ≥10 & *p*adj <0.05 & log2FoldChange >0.4)	80/20 70/30 60/40	No Balancing, RandomUnderSampler, SMOTE, RandomOversampler, TOMEKLinks, BorderlineSMOTE, ADASYN, SVMSMOTE, KMeansSMOTE
DGE (basemean ≥10 & *p*adj <0.05) + ROC Analysis + Genes Validation	80/20 70/30 60/40	No Balancing, RandomUnderSampler, SMOTE, RandomOversampler, TOMEKLinks, BorderlineSMOTE, ADASYN, SVMSMOTE, KMeansSMOTE

DGE: Differential gene expression; padj: *p*-adjusted value; ROC: receiver operating characteristic; SMOTE: synthetic minority over-sampling technique; ADASYN: adaptive synthetic sampling; SVMSMOTE: support vector machine SMOTE.

Kyoto Encyclopedia of Genes and Genomes (KEGG) and GSEA databases to ensure their relevance to clinical PCa pathways (25).

The gene list, initially formatted using Ensembl IDs, was converted to gene symbols using the SynGO online converter (26) After conversion, the gene list was checked against PCa clinical pathways in the GSEA-MSigDB database (27). Genes identified in these clinical pathways were selected for final model construction. This approach, integrating DGE filtering, ROC analysis, and pathway enrichment, ensures that the genes selected are both biologically relevant and predictive, providing a robust foundation for model development.

Data balancing & splitting. After completing the feature selection process, the dataset was found to have a significant class imbalance, with a cancer-to-normal sample ratio of approximately 9:1. To effectively address this imbalance, we first performed stratified train-test splits at ratios of 60/40, 70/30, and 80/20. Stratification ensured that the class distribution was preserved in both training and testing sets, while the test set was kept unchanged to provide a realistic and unbiased evaluation of model performance. Various resampling techniques were then applied exclusively to the training data to rebalance it. These methods included Random Over Sampling (ROS), Random Under Sampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE, TOMEK Links, Adaptive Synthetic Sampling (ADASYN), KMeansSMOTE, and Support Vector Machine SMOTE (SVMSMOTE). The goal was to adjust the class distribution in the training set so that cancer samples constituted approximately 66.66% and normal samples 33.33% of the data, corresponding to a sampling strategy of 0.3. This careful rebalancing aimed to improve the classifier's ability to learn patterns from the minority cancer class while preserving the integrity of the test data for valid model assessment (Table II).

Model construction. The baseline model used in this study is logistic regression, selected for its straightforward approach and interpretability qualities that are particularly valuable when analyzing gene expression data for prostate cancer detection. To benchmark the performance of this baseline model, we compared it with three widely used classifiers: Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). These comparison models were implemented to

evaluate how well logistic regression performs under optimal conditions. Specifically, after identifying the top-performing configuration for logistic regression, we applied the same optimal settings to the other classifiers to ensure a fair performance benchmark. All models were trained and tested using identical preprocessing steps, including normalization and feature selection, to maintain consistency in evaluation.

To ensure the reliability of the trained models, particularly in terms of statistical robustness, the training was performed using data from the White race cohort, which represented the largest population group in the dataset. This choice provides a more stable training foundation, reduces the risk of overfitting due to limited sample sizes, and enhances the generalizability of the findings within the context of the available data.

Additionally, for each given model, Hyperparameter Tuning is applied for improving the model's performance. More details on hyperparameter grid are shown on Table III.

Model evaluation. In this study, we evaluate model performance using a comprehensive classification report that includes Precision, Recall (Sensitivity), Accuracy, and F1-score, supported by a confusion matrix for detailed error analysis. Precision measures the proportion of correct positive predictions, while Recall assesses the model's ability to detect all relevant positive instances. Accuracy reflects the overall correctness of predictions across all classes, and the F1-score provides a balanced harmonic mean of Precision and Recall, accounting for both false positives and false negatives. These metrics collectively help identify issues such as underfitting or overfitting by ensuring no single aspect of performance is overlooked. The confusion matrix further visualizes true positives, true negatives, false positives, and false negatives, offering an intuitive view of classification outcomes. Additionally, we assess the area under the receiver operating characteristic curve (AUC-ROC), which quantifies the model's ability to distinguish between classes across various threshold settings, with higher values indicating better discriminatory power.

Table III. *Hyperparameter grid for Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) classifiers.*

Classifier	Hyperparameters	Values/Options
LR	C	[0.1, 1, 10]
	solver	['liblinear', 'saga', 'lbfgs']
	penalty	['l2']
	max_iter	[1000]
RF	n_estimators	[100, 200]
	max_depth	[None, 10, 20]
	min_samples_split	[2, 5]
	min_samples_leaf	[1, 2]
SVC	C	[0.1, 1, 10]
	kernel	['linear', 'rbf']
	gamma	['scale', 'auto']
KNN	n_neighbors	[3, 5, 7]
	weights	['uniform', 'distance']
	metric	['euclidean', 'manhattan']

To ensure ethical and equitable model deployment, fairness metrics were also incorporated to evaluate whether the model performs consistently across different subgroups or demographics, helping to detect potential biases. Specifically, we assessed Demographic Parity, which evaluates whether positive prediction rates are equally distributed across demographic groups, and Equal Opportunity, which examines whether true positive rates are consistent across those groups. These fairness assessments enable a more responsible evaluation of model outcomes and highlight any disparities that may require mitigation. All performance evaluations were conducted using scikit-learn version 1.5.1. From all tested scenarios, the top five models exhibiting the best overall performance were selected for further benchmarking. These selected models were then evaluated using different dataset cohorts and a variety of classifier algorithms to validate their robustness and generalizability across diverse conditions.

Results and Discussion

Biomarker selection for PCa classification. The feature selection process across the four scenarios resulted in subsets of 4, 9, 19, and 139 as shown in Table IV. These

Table IV. *Gene subsets identified through feature selection methods.*

Feature selection scenario	Identified genes
DGE with basemean ≥10 & *p*adj <0.05 + ROC analysis + MsigDB validation	9
DGE with basemean ≥10 & *p*adj <0.05 & abs(log2FoldChange) >0.4	139
DGE with basemean ≥10 & *p*adj <0.05 & abs(log2FoldChange) >0.35	19
DGE with basemean ≥10 & *p*adj <0.05 & abs(log2FoldChange) >0.4	4

DGE: Differential gene expression; *p*adj: *p*-adjusted value; ROC: receiver operating characteristic; MsigDB: molecular signature database; abs: absolute.

gene subsets identified as the most significant genes for PCa classification. These selected genes were determined based on their relevance to distinguishing between cancerous and non-cancerous samples.

Specifically for the first scenario, we initially obtained a list of 13 filtered genes by using a combination of DGE analysis, ROC analysis, the results of this analysis can be seen in Table V. This gene was then converted to a gene symbol for Gene-Set Enrichment Analysis.

Using the converted gene symbols, a GSEA was performed against the MSigDB database. As shown in Figure 2, the most significant overlap occurred with the LIU_PROSTATE_CANCER_DN gene set (9 overlapping genes out of 493; $p=2.39\times10^{-15}$, FDRq=2.05×10^{-11}), which contains genes down-regulated in PCa samples. This strongly supports the biological relevance of the selected biomarkers to PCa. Importantly, this overlap confirms that the top-ranked genes in our study are not only statistically significant but are also part of a known PCa expression signature, reinforcing their validity as potential diagnostic targets.

Additionally, the gene list showed enrichment in other cancer-related pathways, including DELYS_THYROID_CANCER_DN and SENESE_HDAC1 _AND_ HDAC2_TARGETS_DN, both with FDRq <0.001. These pathways are associated with thyroid carcinoma and epigenetic regulation in osteosarcoma, respectively. Such overlaps suggest that the selected genes may be involved in broader cancer-related regulatory mechanisms, including cellular differentiation, chromatin remodeling, and transcriptional repression. The presence of significant overlaps with multiple cancer modules (*e.g.*, MODULE_11, MODULE_100) and brain tumor gene sets (*e.g.*, JOHANSSON_BRAIN_CANCER_EARLY_VS_LATE) further emphasizes the potential role of these genes in generalized tumor biology.

Table V. *Conversion of Ensembl IDs to gene symbols for selected genes.*

Gene symbol (ENS)	Converted gene symbol
ENSG00000244509.3	*APOBEC3C*
ENSG00000170271.9	*CLU*
ENSG00000084207.14	*CRYAB*
ENSG00000152137.5	*FAM107A*
ENSG00000168077.12	*FAXDC2*
ENSG00000066468.19	*FGFR2*
ENSG00000139926.14	*FRMD6*
ENSG00000109846.6	*GJA1*
ENSG00000134202.9	*GSTM3*
ENSG00000065534.17	*GSTP1*
ENSG00000152661.7	*HSPB8*
ENSG00000120885.18	*MYLK*
ENSG00000168309.15	*SCARA3*

Overall, this analysis not only validates the biological importance of the final 9-gene panel but also highlights their involvement in core oncogenic processes across multiple cancer types. These results strengthen the justification for using these biomarkers in a PCa diagnostic context and support their investigation for broader translational relevance.

Apart from GSEA, we also conducted a statistical analysis using Random Forest, which yielded results consistent with those of GSEA. Among 13 genes, 9 genes were found statistically significant as shown in Figure 3.

The final subset of nine genes *GJA1, FRMD6, FAM107A, HSPB8, GSTP1, MYLK, GSTM3, CRYAB*, and *FGFR2* not only demonstrated statistical significance (adjusted *p*-value <0.05, AUC >0.9), but also show strong biological relevance to prostate cancer pathology. GJA1 (Connexin 43) has been

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDRq-value
LIU_PROSTATE_CANCER_DN [493]	Genes down-regulated in prostate cancer samples.	9		$2.39\ e^{-15}$	$2.05\ e^{-11}$
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_MA_UP [211]	Up-regulated genes in angioimmunoblastic lymphoma (AILT) compared to normal T lymphocytes.	5		$3.55\ e^{-9}$	$1.53\ e^{-5}$
JOHANSSON_BRAIN_CANCER_EARLY_VS_LATE_DE_DN [43]	Genes down-regulated in early vs late brain tumors induced by retroviral delivery of PDGFB [GeneID=5155].	3		$2.72\ e^{-7}$	$4.65\ e^{-4}$
MODULE_11 [540]	Genes in the cancer module 11.	5		$3.81\ e^{-7}$	$4.65\ e^{-4}$
MODULE_100 [544]	Genes in the cancer module 100.	5		$3.95\ e^{-7}$	$4.65\ e^{-4}$
MODULE_137 [546]	CNS genes.	5		$4.03\ e^{-7}$	$4.65\ e^{-4}$
MODULE_66 [552]	Genes in the cancer module 66.	5		$4.25\ e^{-7}$	$4.65\ e^{-4}$
GAVISH_3CA_MALIGNANT_METAPROGRAM_25_AS_ASTROCYTES [50]	Genes upregulated in subsets of cells of a given type within various tumors	3		$4.32\ e^{-7}$	$4.65\ e^{-4}$
DELYS_THYROID_CANCER_DN [233]	Genes down-regulated in papillary thyroid carcinoma (PTC) compared to normal tissue.	4		$6.01\ e^{-7}$	$5.63\ e^{-4}$
SENESE_HDAC1_AND_HDAC2_TARGETS_DN [238]	Genes down-regulated in U2OS cells (osteosarcoma) upon knockdown of both HDAC1 and HDAC2 [GeneID=3065;3066] by RNAi.	4		$6.54\ e^{-7}$	$5.63\ e^{-4}$

Figure 2. *Gene Set Enrichment Analysis (GSEA) using the MSigDB database revealed a significant enrichment of overlapping genes associated with cancer-related clinical pathways, underscoring the biological relevance of the identified gene signatures in the context of tumor progression and diagnosis.*

implicated in tumor suppression and altered expression across multiple cancers, including PCa, due to its role in gap junction communication (29). FRMD6, a regulator of cell polarity, is linked to growth control and favorable outcomes in certain cancers (30). FAM107A has emerged as a tumor suppressor in PCa and is frequently down-regulated in aggressive cases (31). HSPB8, a small heat shock protein, facilitates cancer progression *via* the JAK/STAT3 pathway and is known to be up-regulated in PCa (32). GSTP1, perhaps one of the most established PCa biomarkers, shows differential methylation and expression patterns between racial groups and plays a crucial role in detoxification (33). MYLK has been identified as a predictive marker for PCa recurrence (34), while GSTM3, another detoxification enzyme, has shown polymorphisms associated with prostate cancer risk (35). CRYAB, known

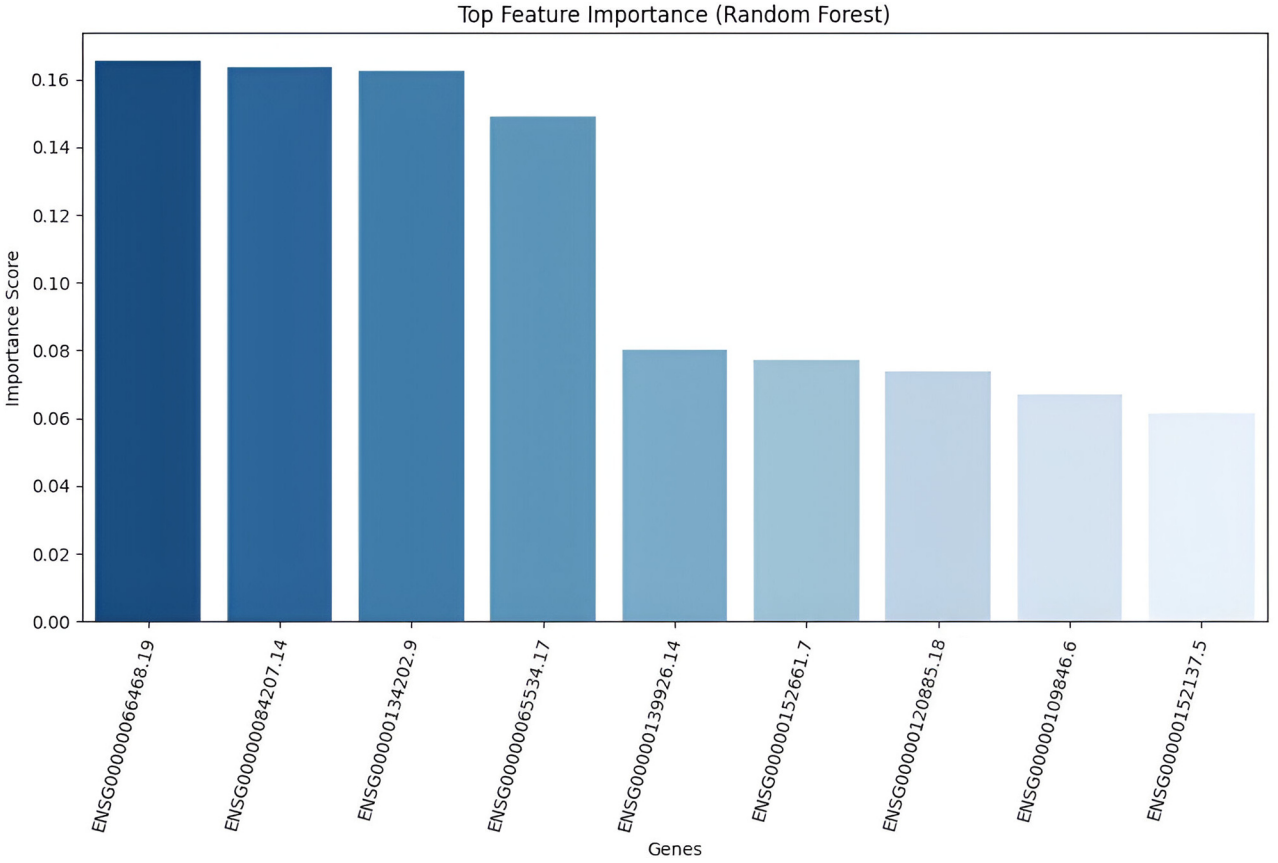

Figure 3. *Top nine genes that are the most important based on statistical analysis.*

for its anti-apoptotic functions, and FGFR2, a regulator of cell proliferation and angiogenesis, are both involved in tumorigenic processes and were found to be significantly altered in prostate malignancies (36, 37). For more details on the overlapped genes, please check Table VI.

Collectively, these genes reflect diverse biological processes including apoptosis regulation, stress response, detoxification, and growth signaling, aligning well with known mechanisms of prostate cancer progression. Their identification through integrated DGE, ROC, and GSEA methods further supports their utility not only in classification tasks but also as potential candidates for biomarker-driven therapeutic interventions.

While our final model included 9 key genes with strong diagnostic performance and enrichment in prostate cancer pathways, several additional genes identified in earlier filtering stages also warrant biological consideration. For instance, Clusterin (CLU) has been

Table VI. *Genes overlapping between enrichment and statistical analyses.*

Overlapped genes from statistical analysis and enrichment analysis	
ENSG00000109846.6	*GJA1*
ENSG00000139926.14	*FRMD6*
ENSG00000152137.5	*FAM107A*
ENSG00000152661.7	*HSPB8*
ENSG00000065534.17	*GSTP1*
ENSG00000120885.18	*MYLK*
ENSG00000134202.9	*GSTM3*
ENSG00000084207.14	*CRYAB*
ENSG00000066468.19	*FGFR2*

Table VII. *Top five logistic regression models in the White population.*

No	Scenario	Balancing technique	Data splitting	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC
1	DGE+ROC+GSEA	SVMSmote	60/40	95	97.1	96.7	97.5	0.96
2	DGE+ROC+GSEA	Random Oversampler	70/30	95	97.5	97.5	97.5	0.95
3	DGE+ROC+GSEA	Random Oversampler	80/20	94.5	95.2	97.5	97	0.92
4	DGE+ROC+GSEA	SMOTE	80/20	94.5	96	95	97.5	0.92
5	DGE+ROC+GSEA	BorderlineSMOTE	80/20	94	97	95.2	98.7	0.92

DGE: Differential gene expression; *p*adj: *p*-adjusted value; ROC: receiver operating characteristic; GSEA: gene set enrichment analysis.

Table VIII. *Top five logistic regression models in the Black population.*

No	Scenario	Balancing technique	Data splitting	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC
1	DGE+ROC+GSEA	RandomUnderSampler	80/20	96.8	96.4	98.2	98.7	0.99
2	DGE+ROC+GSEA	None	80/20	96.8	98.2	98.2	97.5	0.98
3	DGE+ROC+GSEA	SMOTE	80/20	96.8	96.4	98.2	97.5	0.99
4	DGE+ROC+GSEA	TomekLinks	80/20	95.3	94.7	97.2	98.7	0.99
5	DGE+ROC+GSEA	RandomUnderSampler	60/40	92.1	95	92	98	0.97

DGE: Differential gene expression; *p*adj: *p*-adjusted value; ROC: receiver operating characteristic; GSEA: gene set enrichment analysis; AUC: area under curve.

implicated in apoptotic regulation and treatment resistance across several cancers, including prostate tumors. APOBEC3C, part of the APOBEC family of cytidine deaminases, has been associated with somatic mutation processes and may influence tumor heterogeneity. Similarly, SCARA3, although not retained in the final model, plays a role in oxidative stress response and has been shown to modulate tumor cell survival in other malignancies. Although these genes did not meet the final inclusion criteria based on ROC or enrichment thresholds, their presence in early-stage filters suggests potential relevance, and further biological validation could uncover additional roles in PCa progression or resistance mechanisms.

Performance on white population. Using the model construction scenario, there is a significant result in our PCa diagnostic models. The top 5 models are presented in Table VII.

The performance of the model on the White population indicates that the best-performing model, trained using nine genes, achieved an accuracy of 95% and an AUC of 0.96. The high accuracy suggests that the model is highly effective at correctly classifying individuals as either having prostate cancer or not. Meanwhile, the AUC of 0.96 reflects the model's excellent ability to distinguish between cancer and normal samples across all possible classification thresholds indicating strong discriminative power, even when the decision boundary shifts. These genes significantly overlap with the PCa clinical pathways, as demonstrated by GSEA, supporting their biological relevance. This finding suggests that these genes play a crucial role in PCa progression and may serve as key biomarkers for accurate classification. The improved model performance underscores the importance of biologically informed feature selection, where training on disease-specific biomarkers enhances predictive accuracy. Furthermore, this result reinforces the potential of precision medicine approaches, demonstrating that selecting the most relevant molecular features could lead to more effective and tailored diagnostic models for specific populations.

Performance on Black population. Further evaluation of the Black population identified the top five best-performing models, as presented in Table VIII.

When testing the model on the Black population dataset, a similar performance trend was observed, with the best-performing model again derived from the 9-gene subset, achieving an impressive accuracy of 96.8%. Notably, this accuracy surpasses that observed in the White population. This discrepancy may be attributed to the smaller sample size of the Black cohort, which can lead to overrepresentation and potentially inflate performance metrics. Nevertheless, these findings reinforce the earlier results, confirming that this specific set of biomarkers plays a pivotal role in enhancing model accuracy for prostate cancer detection. The consistent performance across racial groups underscores the strong predictive power of these nine genes, positioning them as robust and generalizable biomarkers for prostate cancer classification. Furthermore, this outcome highlights the critical importance of targeted feature selection, demonstrating that training with biologically relevant genes not only improves predictive performance but also reduces dimensionality resulting in a more efficient and interpretable model. The model's best AUC score reached an outstanding 0.99, which is further evidence of its strong discriminative capability.

Performance comparison with alternative classifiers. After identifying the top-performing model scenarios for both populations, we benchmarked these scenarios against alternative classifiers. For the White population, we used SVMSMOTE with nine selected genes and a 60/40 train-test split. For the Black population, we employed RandomUnderSampler with an 80/20 split as the baseline for comparison. The result can be seen in Table IX.

After comparing three different classifiers with our baseline model (logistic regression), the results show that logistic regression still outperforms the others as shown in Table X. Although SVC achieved a similar accuracy, logistic regression yielded a slightly higher AUC score, outperforming SVC. A similar result was observed in the Black population, where logistic regression also outperformed the other classifiers, further reinforcing the robustness of our baseline model.

Table IX. *Classifier performance comparison in the White population.*

No	Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC
1	LR	95.6	97.1	96.7	97.5	0.96
2	SVC	95.6	97.2	97.5	96.9	0.95
3	RF	94.5	96.9	96.4	97.5	0.91
4	KNN	94.5	96.9	96.4	97.5	0.93

LR: Logistic regression; SVC: support vector classifier; RF: random forest; KNN: K-Nearest-Neighbors; AUC: area under curve.

Table X. *Classifier performance comparison in the Black population.*

No	Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC
1	LR	96.8	98.4	98.2	98..7	0.99
2	SVC	93.7	96.4	98.1	94.7	0.98
3	RF	95.3	96.9	96.4	97.5	0.99
4	KNN	94.5	96.5	94.8	98.2	0.98

LR: Logistic regression; SVC: support vector classifier; RF: random forest; KNN: K-Nearest-Neighbors; AUC: area under curve.

Demographic parity assessment. We evaluated the fairness of our top prostate cancer detection model across racial groups using Demographic Parity. The model achieved high detection rates for both White (90%) and Black (86%) cohorts, with a 4% difference, as shown in Figure 4. A chi-square test of independence produced a p-value of 0.518, indicating this difference is not statistically significant (p>0.05) and likely due to random variation rather than algorithmic bias. Although statistically non-significant, the 4% disparity warrants clinical consideration given known prostate cancer outcome differences across races. The model's conservative detection threshold helps minimize false negatives, which are critical in screening settings, and ensures comparable treatment across demographics.

From a fair perspective, the model satisfies Demographic Parity, providing equitable positive prediction rates for White and Black patients when clinical indicators warrant. Observed differences are more likely driven by genuine clinical, genetic, or sociodemographic factors rather than discrimination, supporting the model's responsible clinical use.

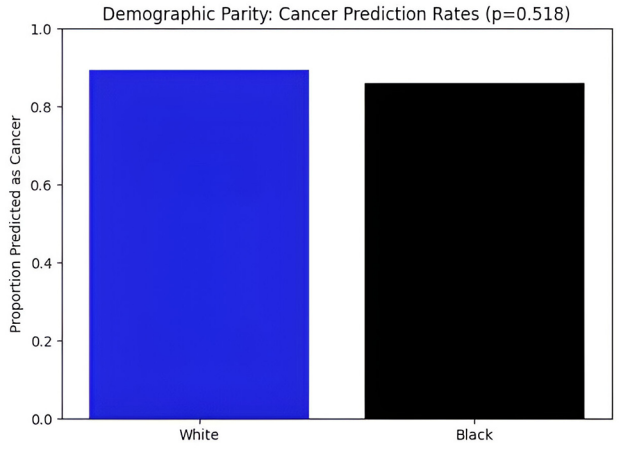

Figure 4. *Demographic parity result across different population datasets.*

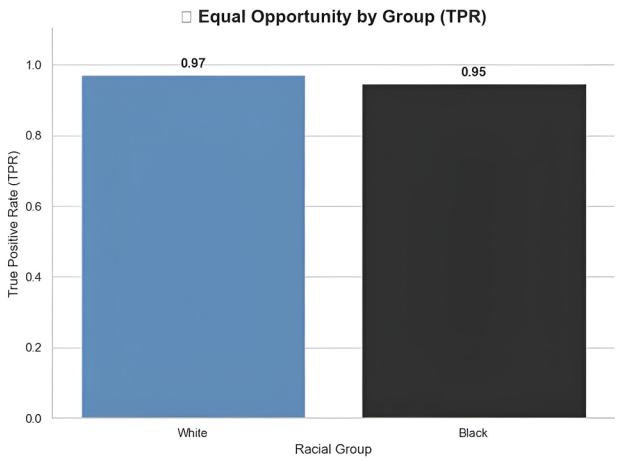

Figure 5. *Equal opportunity assessment results, showing that the true positive rate (TPR) on white cohort is higher than black cohort by 2%.*

Table XI. *Comparative analysis with existing prostate cancer diagnostic models.*

Source	Method	Number of features	Data balancing	Best result
Mohammed *et al.* (16)	Logistic regression	Not using gene	X	Acc: 0.91
Busetto *et al.* (17)	Multiple machine learning model	Not using gene	X	Achieved: 0.77 AUC Score
Erdem and Bozkurt (18)	Multi-layer perceptron	Not using gene	X	Acc: 0.97
Casey *et al.* (19)	RFE	20	√	Acc: 0.96
Bouazza *et al.* (20)	KNN	29	X	Acc: 0.85
Proposed model	Logistic regression – DGE - Race specific	9	√	Acc: 0.95 in White population and Acc: 0.96 in Black population with both AUC higher than 0.95

DGE: Differential gene expression; RFE: recursive feature elimination; KNN: K-Nearest Neighbors; AUC: area under curve; Acc: accuracy.

Equal opportunity assessment. Furthermore, we evaluated Equal Opportunity to assess fairness with respect to true positive rates (TPR) across racial groups. The TPR was 97% for the White cohort and 95% for the Black cohort, yielding a difference of 2%. This indicates that the model's ability to correctly identify cancer cases is relatively balanced across groups. Even though statistically the model is slightly more effective in detecting cancer cases in white cohort, the gap is small (2%), which is generally acceptable (Figure 5).

Comparison with existing models. To demonstrate the effectiveness of our approach, we conducted a comparative analysis with relevant existing research, as presented in Table XI. While most prior studies rely on imaging data, those utilizing gene expression typically involve larger gene sets, ranging from 20 to 29 genes. In contrast, our proposed model achieves comparable or even superior diagnostic accuracy using only nine genes, emphasizing its efficiency and suitability for clinical application. Moreover, our framework uniquely accounts for racial disparities and incorporates fairness metrics, ensuring both high performance and equitable outcomes across diverse populations. Our research further strengthens its impact through rigorous biomarker selection and the integration of race-aware modeling, which together enhance both the diagnostic precision and the fairness of the model in real-world clinical settings.

Limitations & future works. While this study demonstrates promising results in prostate cancer (PCa) diagnosis, several limitations must be acknowledged. Despite employing a thorough feature selection method, the dataset suffers from significant class imbalance, with 414 cancer samples compared to only 44 normal samples, a roughly 9:1 ratio. This heavy imbalance may adversely affect the model's accuracy, particularly for the underrepresented normal class. Additionally, the cross-validation performed on the Black population subset is limited by a small sample size of only 64 individuals, which may lead to overly optimistic results due to easier data separability and reduced generalizability. Another limitation is the reliance on secondary data from the TCGA repository, which restricts the scope and diversity of experiments. The diagnostic model could be substantially improved by incorporating primary gene expression data collected directly from clinical hospital sources, reflecting more varied and real-world conditions. Furthermore, computational constraints limit extensive hyperparameter tuning and the exploration of alternative optimization methods, which could enhance model performance and robustness. Addressing these limitations in future work will be essential to improve the reliability and clinical applicability of prostate cancer diagnostic tools, especially across diverse racial groups.

Conclusion

In this study, we propose a novel prostate cancer detection framework that accounts for racial disparities and leverages targeted biomarkers to enhance diagnostic accuracy. Unlike traditional models that assume a one-size-fits-all approach, our framework integrates race-specific gene expression patterns, enabling the model to capture biological variations across populations more effectively. Our findings reveal that racial disparities in cancer detection are a significant factor, with up to a 4% difference in diagnostic performance between racial groups. This highlights the value of developing more inclusive, population-aware approaches in biomedical artificial intelligence to improve equity and outcomes across diverse communities. While our results are promising, we recognize the opportunity to further strengthen the framework by expanding the dataset particularly for underrepresented groups such as the Black cohort. This will help ensure even greater robustness and generalizability. Nonetheless, our study offers a strong foundation and compelling evidence that personalized, race-aware models can play a pivotal role in advancing precision diagnostics and addressing healthcare disparities.

Conflicts of Interest

The Authors declare no conflicts of interest in relation to this study.

Authors' Contributions

Conceptualization: DA (lead), MVO (equal), MW (equal), VK (equal); Data curation: VK (lead), AM (supporting), JS (supporting); Formal analysis: VK; Funding acquisition: DA; Investigation: VK; Methodology: DA (lead), MVO (equal), MW (equal), VK (equal), MIA (equal); Project administration: VK (lead), AM (supporting), JS (supporting); Resources: VK (lead), AM (supporting), JS (supporting); Supervision: DA (lead), MVO (equal), MW (equal), MIA (equal); Validation: DA (lead), MVO (equal), MW (equal), MIA (equal); Visualization: VK (lead), AM (equal), JS (supporting); Writing – original draft: VK (lead), AM (supporting), JS (supporting); Writing – review & editing: VK (lead), AM (supporting), JS (supporting).

Acknowledgements

The Authors gratefully acknowledge Universitas Multimedia Nusantara for their invaluable support throughout this study. The resources, guidance, and encouragement provided played a vital role in the successful completion of this research.

Funding

This research was funded by Kemerinstekdikti Indonesia under contract number 105/E5/PG.02.00.PL/2024 and decree number 0459/E5/PG.02.00/2024.

Artificial Intelligence (AI) Disclosure

During the preparation of this manuscript, a large language model (ChatGPT, OpenAI) was used solely for language editing and stylistic improvements in select paragraphs. No sections involving the generation, analysis, or interpretation of research data were produced by generative AI. All scientific content was created and verified by the authors. Furthermore, no figures or visual data were generated or modified using generative AI or machine learning-based image enhancement tools.

References

1 Vietri MT, D'Elia G, Caliendo G, Resse M, Casamassimi A, Passariello L, Albanese L, Cioffi M, Molinari AM: Hereditary prostate cancer: genes related, target therapy and prevention. Int J Mol Sci 22(7): 3753, 2021. DOI: 10.3390/ijms22073753

2 Rawla P: Epidemiology of prostate cancer. World J Oncol 10(2): 63-89, 2019. DOI: 10.14740/wjon1191

3 DeSantis CE, Miller KD, Goding Sauer A, Jemal A, Siegel RL: Cancer statistics for African Americans, 2019. CA Cancer J Clin 69(3): 211-233, 2019. DOI: 10.3322/caac.21555

4 Mahal BA, Gerke T, Awasthi S, Soule HR, Simons JW, Miyahira A, Halabi S, George D, Platz EA, Mucci L, Yamoah K: Prostate cancer racial disparities: a systematic review by the Prostate Cancer Foundation panel. Eur Urol Oncol 5(1): 18-29, 2022. DOI: 10.1016/j.euo.2021.07.006

5 Lumbreras B, Parker LA, Caballero-Romeu JP, Gómez-Pérez L, Puig-García M, López-Garrigós M, García N, Hernández-Aguado I: Variables associated with false-positive PSA results: a cohort study with real-world data. Cancers (Basel) 15(1): 261, 2022. DOI: 10.3390/cancers15010261

6 Srivastava S, Koay EJ, Borowsky AD, De Marzo AM, Ghosh S, Wagner PD, Kramer BS: Cancer overdiagnosis: a biological challenge and clinical dilemma. Nat Rev Cancer 19(6): 349-358, 2019. DOI: 10.1038/s41568-019-0142-8

7 Verbeek JFM, Roobol MJ, ERSPC Rotterdam study group: What is an acceptable false negative rate in the detection of prostate cancer? Transl Androl Urol 7(1): 54-60, 2018. DOI: 10.21037/tau.2017.12.12

8 Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED, Crowley JJ, Coltman CA Jr: Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter. N Engl J Med 350(22): 2239-2246, 2004. DOI: 10.1056/NEJMoa031918

9 Yaqoob A, Musheer Aziz R, Verma NK: Applications and techniques of machine learning in cancer classification: a systematic review. Hum-Cent Intell Syst 3(4): 588-615, 2023. DOI: 10.1007/s44230-023-00041-3

10 Chen S, Jian T, Chi C, Liang Y, Liang X, Yu Y, Jiang F, Lu J: Machine learning-based models enhance the prediction of prostate cancer. Front Oncol 12: 941349, 2022. DOI: 10.3389/fonc.2022.941349

11 Nitta S, Tsutsumi M, Sakka S, Endo T, Hashimoto K, Hasegawa M, Hayashi T, Kawai K, Nishiyama H: Machine learning methods can more efficiently predict prostate cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. Prostate Int 7(3): 114-118, 2019. DOI: 10.1016/j.prnil.2019.01.001

12 Tao J, Bian X, Zhou J, Zhang M: From microscopes to molecules: The evolution of prostate cancer diagnostics. Cytojournal 21: 29, 2024. DOI: 10.25259/Cytojournal_36_2024

13 Human genomic variation. Genome.gov, February 1, 2023. Available at: https://www.genome.gov/ [Last accessed on March 13, 2025]

14 Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS, Al Harbi S, Albekairy AM: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 23(1): 689, 2023. DOI: 10.1186/s12909-023-04698-z

15 Ho D, Quake SR, McCabe ERB, Chng WJ, Chow EK, Ding X, Gelb BD, Ginsburg GS, Hassenstab J, Ho CM, Mobley WC, Nolan GP, Rosen ST, Tan P, Yen Y, Zarrinpar A: Enabling technologies for personalized and precision medicine. Trends Biotechnol 38(5): 497-518, 2020. DOI: 10.1016/j.tibtech.2019.12.021

16 Mohammed Ismail B, Alam M, Tahernezhadi M, Vege HK, Rajesh P: A machine learning classification technique for predicting prostate cancer. 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA, 228-232, 2020. DOI: 10.1109/EIT48999.2020.9208240

17 Busetto GM, Del Giudice F, Maggi M, De Marco F, Porreca A, Sperduti I, Magliocca FM, Salciccia S, Chung BI, De Berardinis E, Sciarra A: Prospective assessment of two-gene urinary test with multiparametric magnetic resonance imaging of the prostate for men undergoing primary prostate biopsy. World J Urol 39(6): 1869-1877, 2021. DOI: 10.1007/s00345-020-03359-w

18 Erdem E, Bozkurt F: A comparison of various supervised machine learning techniques for prostate cancer prediction. Avrupa Bilim Teknol Derg 21: 610-620, 2021. DOI: 10.31590/ejosat.802810

19 Casey M, Chen B, Zhou J, Zhou N: A machine learning approach to prostate cancer risk classification through use of RNA sequencing data. Lecture Notes in Computer Science: 65-79, 2019. DOI: 10.1007/978-3-030-23551-2_5

20 Bouazza SH, Hamdi N, Zeroual A, Auhmani K: Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. 2015 Intelligent

Systems and Computer Vision (ISCV), Fez, Morocco, pp. 1-6, 2015. DOI: 10.1109/ISACV.2015.7106168
21 Kwabi-Addo B: Epigenetic biomarkers and racial differences in cancer. Epigenetic Mechanisms in Cancer: 243-273, 2018. DOI: 10.1016/B978-0-12-809552-2.00010-3
22 Wright CA, Gordon ER, Cooper SJ: Genomic analysis reveals HDAC1 regulates clinically relevant transcriptional programs in Pancreatic cancer. BMC Cancer 23(1): 1137, 2023. DOI: 10.1186/s12885-023-11645-0
23 Obuchowski NA, Lieber ML, Wians FH Jr: ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 50(7): 1118-1125, 2004. DOI: 10.1373/clinchem.2004.031823
24 Yang B, Jiang Y, Yang J, Zhou W, Yang T, Zhang R, Xu J, Guo H: Characterization of metabolism-associated molecular patterns in prostate cancer. BMC Urol 23(1): 104, 2023. DOI: 10.1186/s12894-023-01275-w
25 Tihagam RD, Bhatnagar S: A multi-platform normalization method for meta-analysis of gene expression data. Methods 217: 43-48, 2023. DOI: 10.1016/j.ymeth.2023.06.012
26 SyngoPortal. Available at: https://syngoportal.org/ [Last accessed on March 18, 2025]
27 GSEA Molecular Signatures Database (MSigDB). Available at: https://www.gsea-msigdb.org/gsea/msigdb/ [Last accessed on March 18, 2025]
28 James G, Witten D, Hastie T, Tibshirani R: An introduction to statistical learning. 1st ed. Springer Texts in Statistics. New York, NY, USA, Springer, 2013.
29 Xu H, Wang X, Zhu F, Guo S, Chao Z, Cao C, Lu Z, Zhu H, Wang M, Zhu F, Yang J, Zeng R, Yao Y: Comprehensive pan-cancer analysis of connexin 43 as a potential biomarker and therapeutic target in human kidney renal clear cell carcinoma (KIRC). Medicina (Kaunas) 60(5): 780, 2024. DOI: 10.3390/medicina60050780
30 von Koskull A, Hagström J, Haglund C, Kaprio T, Böckelman C: High-tissue FRMD6 expression predicts better outcomes among colorectal cancer patients. Biomarkers 29(3): 127-133, 2024. DOI: 10.1080/1354750X.2024.2321916
31 Ma YF, Li GD, Sun X, Li XX, Gao Y, Gao C, Cao KX, Yang GW, Yu MW, Wang XM: Identification of FAM107A as a potential biomarker and therapeutic target for prostate carcinoma. Am J Transl Res 13(9): 10163-10177, 2021.
32 Zhang K, Yin W, Ma L, Liu Z, Li Q: HSPB8 facilitates prostate cancer progression *via* activating the JAK/STAT3 signaling pathway. Biochem Cell Biol 101(1): 1-11, 2023. DOI: 10.1139/bcb-2022-0205
33 Vidal I, Zheng Q, Hicks JL, Chen J, Platz EA, Trock BJ, Kulac I, Baena-Del Valle JA, Sfanos KS, Ernst S, Jones T, Maynard JP, Glavaris SA, Nelson WG, Yegnasubramanian S, De Marzo AM: GSTP1 positive prostatic adenocarcinomas are more common in Black than White men in the United States. PLoS One 16(6): e0241934, 2021. DOI: 10.1371/journal.pone.0241934
34 Qiao P, Zhang D, Zeng S, Wang Y, Wang B, Hu X: Using machine learning method to identify *MYLK* as a novel marker to predict biochemical recurrence in prostate cancer. Biomark Med 15(1): 29-41, 2021. DOI: 10.2217/bmm-2020-0495
35 Wang S, Yang J, You L, Dai M, Zhao Y: GSTM3 function and polymorphism in cancer: emerging but promising. Cancer Manag Res 12: 10377-10388, 2020. DOI: 10.2147/CMAR.S272467
36 Zhang CL, Hu Y, Wang DX, Yang Q, Chang DH: [Research on the anti-tumor effects of CRYAB in prostate cancer]. Zhonghua Nan Ke Xue 29(7): 579-586, 2023.
37 Lee JE, Shin SH, Shin HW, Chun YS, Park JW: Nuclear FGFR2 negatively regulates hypoxia-induced cell invasion in prostate cancer by interacting with HIF-1 and HIF-2. Sci Rep 9(1): 3480, 2019. DOI: 10.1038/s41598-019-39843-6