CANCER DIAGNOSIS & PROGNOSIS *6*: 660-673 (2026)
doi: 10.21873/cdp.10566

Optimized Prostate Cancer Stage Classification Using XGBoost Based on Racial miRNA Expression

JHENO SYECHLO[1], DAVID AGUSTRIAWAN[1], VINCENT KURNIAWAN[1], ADITHAMA MULIA[1], EZRA BERNARDUS WIJAYA[2], RIZKY NURDIANSYAH[3], DINAR AJENG KRISTIYANTI[1] and AJIE KUSUMA WARDHANA[1]

[1]Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia;
[2]Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan, R.O.C.;
[3]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, QLD, Australia

Abstract

Background/Aim: Prostate cancer has a high mortality rate and shows diagnostic disparities between racial groups, particularly between White and Black populations. This study aimed to develop a prostate cancer stage classification model using the XGBoost algorithm on miRNA expression data with a focus on race-based analysis.

Materials and Methods: The data used in this study were obtained from the Genomic Data Commons. The Cancer Genome Atlas through the UCSC Xena Browser, consisting of miRNA expression and patient clinical data. Several feature selection methods were applied, including Student's *t*-test, mutual information, chi-square, and limma. Data balancing techniques such as RandomOversampler, SMOTE, SMOTEENN, and BorderlineSMOTE were also implemented to address class imbalance.

Results: The results showed that the XGBoost model achieved an accuracy of up to 89% on data from White patients. Additionally, the F1-score shows the results of 92%, these results suggest this model is particularly strong in identifying minority class in this unbalance data. However, when tested on data from Black patients, the accuracy decreased to 72-74%, indicating limitations in cross-race performance. Additionally, Local Interpretable Model-agnostic Explanations (LIME) was implemented to identify the gene features that contributed most significantly to the model's predictions.

Conclusion: The study demonstrates that XGBoost, combined with race-specific feature selection and hyperparameter tuning, can accurately detect prostate cancer with high performance. These findings highlight the potential of XGBoost for improving prostate cancer detection while also emphasizing the importance of considering race-specific differences in machine learning models.

Keywords: Feature selection, miRNA, prostate cancer, racial disparity, XGBoost.

David Agustriawan, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Jl. Scientia Boulevard, Gading Serpong, Kel. Curug Sangereng, Kec. Kelapa Dua, Kabupaten Tangerang, Banten, 15810 Indonesia. Tel: +62 877 8153 5936, e-mail: david.agustriawan@umn.ac.id

Received January 14, 2026 | Revised March 8, 2026 | Accepted April 15, 2026

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
©2026 The Author(s). Published by the International Institute of Anticancer Research.

Introduction

Prostate cancer is the second most commonly diagnosed cancer among men worldwide, with an estimated 1,414,000 new cancer cases and 375,304 deaths in 2020 (1). It is the most frequently diagnosed cancer in 112 countries, and the leading cause of cancer death in 48 countries (2). Currently, machine learning techniques have been used in research to further develop the diagnosis of prostate cancer (3, 4). Machine learning techniques can discover patterns from complex datasets that could effectively predict the outcome of prostate cancer (5).

XGBoost, also known as eXtreme Gradient Boosting, is one of many machine learning algorithms that have a more advanced implementation than gradient boosting (6, 7). The algorithm is called "Extreme" because it uses regularization to prevent overfitting (8, 9). Unlike the LSTM algorithm, XGBoost is not an artificial neural network (ANN) but rather an ensemble of decision trees. XGBoost is one of the most commonly used methods for building predictive models due to its accuracy, efficiency, and adaptability to various datasets (10-12). Additionally, the XGBoost algorithm can be used for binary classification, which helps in achieving accurate model predictions. XGBoost can classify miRNA data into early-stage and late-stage classes. In several cases, Ogunleye and Wang demonstrated the strong performance of XGBoost in liver disease, achieving high accuracy and sensitivity (13). This makes XGBoost a robust tool for prostate cancer detection. While XGBoost algorithms can enhance the accuracy of prostate cancer, the outcome is also affected by other key factors. One key factor is racial disparity in prostate cancer (14).

MicroRNA (miRNA) is a group of small RNA molecules measuring 19-25 nucleotides in length. A single miRNA can influence the expression of other miRNAs that are often involved in functional interaction pathways (15, 16). miRNAs control various biological processes such as cell division, cell differentiation, angiogenesis, migration, apoptosis, and oncogenesis (17-19). Dysregulation of miRNA expression in cancer cells is often rooted in the genomic location that encodes the miRNA. They are frequently located in genetically unstable regions, fragile sites, or cancer-associated genomic regions (CAGR), which often leads to their deletion, resulting in a lack of miRNA expression (20). Other than that, each miRNA can have multiple targeted genes (21, 22). This broad targeting allows miRNAs to regulate complex biological pathways. miRNAs are connected to the central dogma of molecular biology, which describes the transfer of genetic information from DNA to RNA and ultimately to proteins. In this process, miRNAs play a role in regulating gene expression by interacting with messenger RNA (mRNA) at the post-transcriptional stage (23). In the context of the central dogma, miRNA is a transcription product of DNA that does not undergo translation into protein but instead acts as a regulator that controls cell growth in the body. Uncontrolled cell growth can lead to the development of cancer cells within the body (24, 25).

Racial differences play a significant role in the diagnosis of prostate cancer, affecting detection outcomes across different racial groups. In the United States, Black men are 1.76 times more likely to be diagnosed with prostate cancer and have a 2.14 times higher mortality rate compared to White men (26). Furthermore, Black men are more likely to be diagnosed at a more advanced stage of the disease (27). These disparities highlight the need for further race-specific analysis to reduce the risk of misdiagnosis and improve detection accuracy. Therefore, understanding racial differences is essential to ensure more accurate prostate cancer diagnosis.

To date, only a small number of studies have explored the use of the XGBoost algorithm for the analysis of miRNA data. A study conducted by Kalaiyarasi *et al.* used a miRNA dataset on prostate cancer. This study performed feature selection from a website to identify the most significant miRNAs, resulting in 50 miRNAs. The study achieved an accuracy of 91% and an AUC of 94% using the SVM algorithm (28). Another study conducted by Fernando *et al.* used 4 miRNAs specifically without feature selection to get the accuracy for biomarkers in prostate cancer. The study achieved accuracy of 85% by using logistic regression (29).

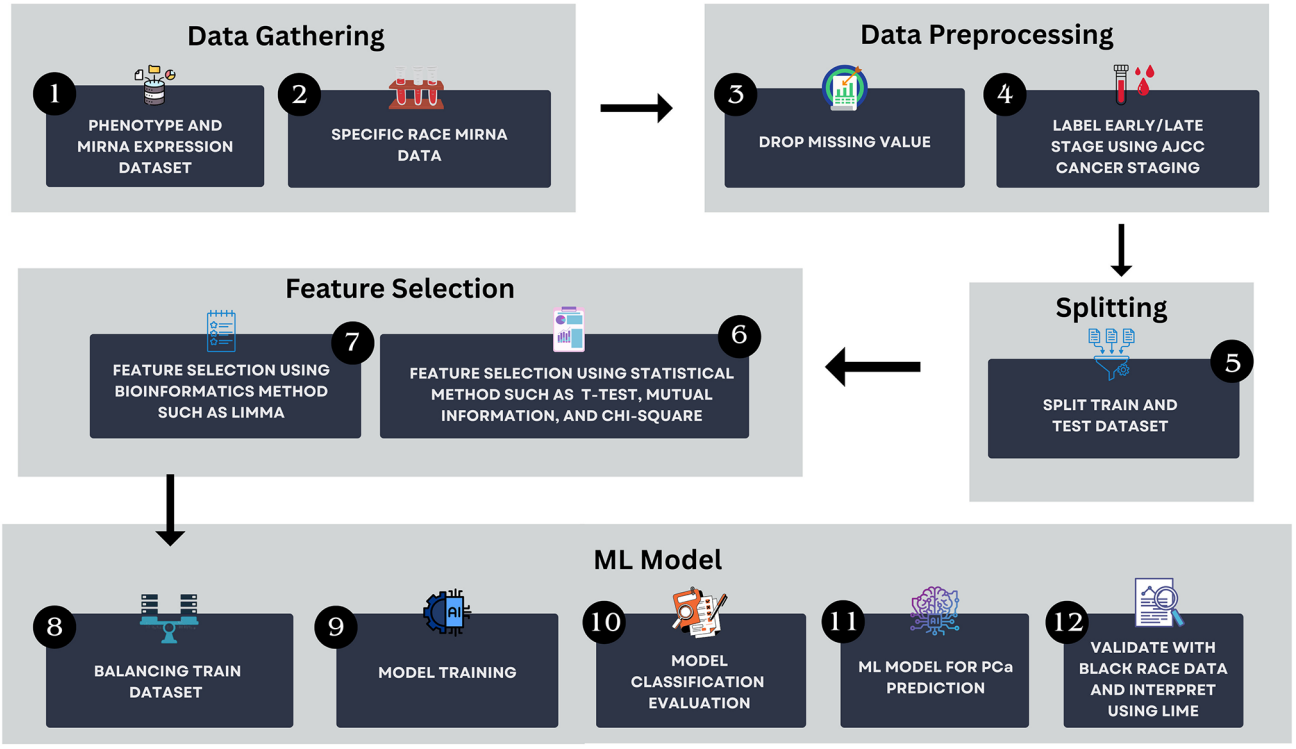

Figure 1. *Research pipeline for prostate cancer classification using miRNA expression data, including data collection, preprocessing, feature selection, model training, evaluation, prediction, and validation with local interpretable model-agnostic explanations (LIME) interpretation.*

Unlike previous research, the current study used data obtained from the GDC TCGA XenaBrowser and applied the XGBoost algorithm, which has not yet been used in prostate cancer staging detection. Moreover, the number of miRNA feature that has been used in this study from feature selection technique has lower number. Furthermore, this study employs different feature selection methods to identify significant miRNA features, such as Student's *t*-test, mutual information, chi-square and limma. The study used various balancing techniques to achieve balanced data. The data used focuses on patients of the white race, and testing was conducted on black race patients to observe the impact of racial differences. Additionally, the model was interpreted using LIME to check the feature importance for classification. By using a combination of different algorithms and feature selection methods, this research aimed to build a more accurate classification model for prostate cancer staging, allowing patients to detect prostate cancer at an earlier stage.

Materials and Methods

The pipeline of this study is described in Figure 1. The method consists of several steps such as data gathering, data preprocessing, feature selection and AI modelling using XGBoost (v. 3.0.0) All steps in this study were conducted using python version 3.13.2 in visual studio software version 1.109 (Microsoft Corporation, Redmond, WA, USA). The devices used for this study included a Windows 11 OS, with 16GB RAM, a 12th Gen Intel® Core™ i5-12500H (16 cores) processor, and NVIDIA GeForce RTX 3050 4GB.

Data gathering. This study utilized two primary datasets: stem-loop miRNA expression data and GDC TCGA-PRAD

phenotype data, both obtained from the GDC TCGA Xena Browser on February 3, 2025 (30). The stem-loop miRNA expression dataset contains miRNA expression levels from prostate cancer patients, which are essential for identifying disease-specific patterns. The expression values were pre-normalized using the log2(RPM+1) transformation. Additionally, the GDC TCGA-PRAD phenotype dataset includes clinical information such as T stage, M stage, PSA levels, Gleason scores, age, and other relevant clinical variables.

Data preprocessing. In the data preprocessing stage, the miRNA expression dataset was separated based on race using the phenotype dataset as a reference. The separated data then underwent a cleaning process to handle missing values. After the initial preprocessing steps, labelling was carried out for cancer stages 1 through 4, which were subsequently grouped into early-stage and late-stage categories. The labelling was based on T, N, M values, Gleason score, and PSA levels. T values indicate the size and spread of the primary tumour, N values indicate the involvement of nearby lymph nodes, M values indicate the present of distant metastasis, while Gleason score was used to assess the aggressiveness of prostate cancer and PSA levels refer to the measurement of prostate-specific antigen in the blood. The rules used for label grouping are detailed in Table I (31). These criteria, T, N, M values, Gleason score, and PSA, were obtained from the phenotype dataset. Libraries such as Pandas version 2.2.2 and Numpy version 1.26.4 were used for data preprocessing, while Scikit-learn version 1.5.1 was used to perform the labelling process.

Feature selection. After obtaining data labelled as early and late stage, feature selection was conducted to identify the most significant and cancer-related miRNA features. Feature selection was conducted to identify significant miRNA features associated with prostate cancer stage classification. In this study, statistical and information-based feature selection methods were applied, *i.e.* Student's *t*-test, mutual Information, and chi-square, by specifying the desired number of features to be selected from the dataset. The Student's *t*-test was used to determine whether the mean expression levels of miRNAs differed significantly between classes. Mutual Information was used to measure the dependency between miRNA features and the class labels, while the chi-square test evaluated the statistical association between miRNA features and the target variable (32).

In addition, a bioinformatics-based differential expression analysis method, limma, was used to identify significant miRNA features based on *p*-Values obtained from linear modelling and empirical Bayes moderation (33). The feature selection methods used to identify significant miRNA features are summarized in Table II. Scikit-learn version 1.5.1 was used to implement the Mutual Information and chi-square methods, while limma was implemented using R version 4.0.16.

XGBoost modelling. In the model development phase, the data was split into training and testing sets using three different ratios: 80/20, 70/30, and 60/40. To address class imbalance, a balancing technique with a ratio of 1:3 was applied. Only the training data was balanced to avoid introducing synthetic data into the test set. The balancing techniques used in this study include RandomOverSampler, SMOTE, SMOTEENN, and BorderlineSMOTE. Every balancing technique mentioned has a different way to balance the data. RandomOversampler duplicates the minority class to increase its sample based on a selected ratio (34). On the other hand, SMOTE generates synthetic samples for the minority class based on the selected scenario (35). SMOTEENN is similar to SMOTE but uses Edited Nearest Neighbours to get synthetic samples, while BorderlineSMOTE creates synthetic samples based on samples located near the decision boundary to generate new synthetic samples (36, 37). The package used for data balancing was imbalanced-learn version 0.12.3.

The model involved several scenarios focusing on feature selection methods, including Student's *t*-test, mutual information, chi-square and limma.

Table I. *Prostate cancer staging criteria.*

Group	T	N	M	PSA	Gleason
I	T1a-c	N0	M0	PSA <10	Gleason ≤6
	T2a	N0	M0	PSA <10	Gleason ≤6
	T1-2a	N0	M0	PSA X	Gleason X
IIA	T1a-c	N0	M0	PSA <20	Gleason 7
	T1a-c	N0	M0	PSA ≥10<20	Gleason ≤6
	T2a	N0	M0	PSA ≥10<20	Gleason ≤6
	T2a	N0	M0	PSA <20	Gleason 7
	T2b	N0	M0	PSA <20	Gleason ≤7
	T2b	N0	M0	PSA X	Gleason X
IIB	T2c	N0	M0	Any PSA	Any Gleason
	T1-2	N0	M0	PSA ≥20	Any Gleason
	T1-2	N0	M0	Any PSA	Gleason ≥8
III	T3a-b	N0	M0	Any PSA	Any Gleason
IV	T4	N0	M0	Any PSA	Any Gleason
	Any T	N1	M0	Any PSA	Any Gleason
	Any T	Any N	M1	Any PSA	Any Gleason

PSA: Prostate-specific antigen.

Table II. *Feature selection methods.*

No.	Feature Selection Method	Criteria
1	limma	*p*-Value <0.01
2	Student's *t*-test	Best result from 15, 16, 17, 18, and 19 miRNA features
3	Mutual information	Best result from 15, 16, 17, 18, and 19 miRNA features
4	Chi-square	Best result from 15, 16, 17, 18, and 19 miRNA features

Table III. *Modelling scenarios using XGBoost.*

Feature Selection	Splitting Ratio	Balancing Technique
limma criteria *p*-Value <0.01	80:20, 70:30, 60:40	RandomOversampler, SMOTEENN, SMOTE, Borderline SMOTE
T-test the best result from 15, 16, 17, 18 and 19 feature	80:20, 70:30, 60:40	RandomOversampler, SMOTEENN, SMOTE, Borderline SMOTE
Mutual information the best result from 15, 16, 17, 18 and 19 feature	80:20, 70:30, 60:40	RandomOversampler, SMOTEENN, SMOTE, Borderline SMOTE
Chi-square the best result from 15, 16, 17, 18 and 19 feature	80:20, 70:30, 60:40	RandomOversampler, SMOTEENN, SMOTE, Borderline SMOTE

Hyperparameter tuning was performed using Grid Search with cross-validation to obtain optimal results. Hyperparameters such as max_depth and learning_rate were also used during model construction. The scenarios used in the model development process are presented in Table III.

The resulting model was evaluated using a classification report and confusion matrix, which included accuracy, precision, recall, and F1-score metrics. Each metric serves a specific function and is essential in determining whether the model's performance meets the desired criteria. The finalized model was then tested using

a different dataset, specifically, data from Black patients to assess the significance of racial differences. The dataset for Black patients consisted of 58 individuals, providing a representative subset to examine potential biases in classification outcomes.

Results

Data gathering and preprocessing. The datasets used in this study are the stem-loop miRNA expression and GDC TCGA-PRAD phenotype datasets. The stem-loop miRNA expression dataset contains 1,882 miRNA features across 551 samples, as shown in Table IV. Meanwhile, the GDC TCGA-PRAD phenotype dataset includes 88 variables related to clinical and hospital data. This study focuses on the White race, which comprises approximately 83% (458 out of 551) of the total sample population.

Data preprocessing involved labeling each sample using the AJCC Cancer Staging Manual, which provides information on prostate cancer stages from stage 1 to stage 4. Out of 458 patients, 360 patients had identifiable cancer stages. From this group, 309 were classified as early-stage prostate cancer and 51 as late-stage prostate cancer. The resulting labels were stored in the metadata to facilitate the modelling process.

Feature selection. Feature selection was performed using 5 different approaches: Student's *t*-test, mutual information, chi-square and limma. The limma method resulted in a total of 20 selected miRNAs. The criteria used in this scenario were p-value <0.01. For machine learning-based feature selection, Student's *t*-test, mutual information and chi-square test were implemented. First, feature selection was performed using the Student's *t*-test (f_classif) implemented in Scikit-learn, with feature standardization applied using StandardScaler. Next, mutual information (mutual_info_classif) was used to measure the dependency between miRNA features and the class labels, also using standardized features. In addition, the chi-square method was applied after feature normalization using MinMaxScaler, as this method requires non-negative feature values. The SelectKBest approach was used to specify the number of selected features (k) generating candidate miRNA feature subsets consisting of 15, 16, 17, 18, and 19 miRNAs for model training. Using the SelectKBest approach, the number of selected features was determined by specifying the desired number of features (k). The resulting miRNA feature subsets showed significant expression differences and were used for subsequent model training. The results of feature selection are presented in Table V. The selected features represent the most optimal set for model construction. Additionally, feature selection happens after splitting to prevent data leakage while training model for white patient.

XGBoost modelling. In model development, several scenarios were implemented based on the training–test ratio, the number of miRNAs, the feature selection methods used, and the balancing techniques applied. As shown in Table VI, the best performance was achieved when using the Student's *t*-test feature selection method with 15 miRNA features and an 80/20 training–test ratio, where both RandomOverSampler and SMOTE produced the highest test accuracy of 83% with a training accuracy of 91%. In this configuration, the model achieved a precision of 86%, recall of 97%, and an F1-score of 91%, indicating strong classification performance on minority class. Additionally, the results showed relatively consistent performance across different data split ratios, including 80/20, 70/30, and 60/40, where test accuracy ranged from 81% to 83%, suggesting that variations in the training–test ratio and balancing techniques such as RandomOverSampler, SMOTE, and BorderlineSMOTE did not significantly affect the overall model performance. However, when the model was evaluated using internal data from Black patients, the accuracy decreased to 72%, indicating that a model primarily trained on data from White patients may have limited generalizability across different racial groups.

Various experimental scenarios were evaluated by varying the number of miRNAs, feature selection methods,

Table IV. *Expression levels of selected miRNAs across representative TCGA prostate cancer samples.*

miRNA_ID	TCGA-KK-A8IG-01A	TCGA-EJ-7792-11A	TCGA-HC-7079-01A
hsa-let-7a-1	13.42	12.28	12.51
hsa-let-7a-2	13.41	12.27	12.53
hsa-let-7a-3	13.41	12.28	12.52

Table V. *Feature selection results.*

No.	Feature Selection Method	Criteria	miRNA Feature
1	limma	*p*-Value <0.01	20 miRNA
2	Student's *t*-test	Best result from 15, 16, 17, 18, 19 miRNA features	15, 16, 17, 18, 19 miRNA
3	Mutual information	Best result from 15, 16, 17, 18, 19 miRNA features	15, 16, 17, 18, 19 miRNA
4	Chi-square	Best result from 15, 16, 17, 18, 19 miRNA features	15, 16, 17, 18, 19 miRNA

Table VI. *Results for using 15 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
15	Student's *t*-test	80/20	RandomOverSampler	91%	83%	86%	97%	91%	72%
15	Student's *t*-test	80/20	SMOTE	91%	83%	86%	97%	91%	72%
15	Student's *t*-test	70/30	RandomOverSampler	89%	82%	86%	95%	90%	72%
15	Mutual Information	60/40	SMOTE	83%	81%	86%	94%	90%	72%
15	Mutual Information	60/40	BorderlineSMOTE	83%	81%	86%	94%	90%	72%

Acc: Accuracy; Prec: precision.

Table VII. *Results after using 16 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
16	Mutual Information	60/40	BorderlineSMOTE	81%	85%	86%	98%	92%	72%
16	Mutual Information	60/40	SMOTE	83%	83%	86%	95%	90%	72%
16	Mutual Information	60/40	SMOTEENN	96%	82%	86%	95%	90%	72%
16	Student's *t*-test	80/20	RandomOverSampler	92%	82%	86%	95%	90%	72%
16	Student's *t*-test	70/30	SMOTEENN	98%	81%	85%	94%	89%	72%

Acc: Accuracy; Prec: precision.

training–test split ratios, and balancing techniques. As shown in Table VII, the highest performance was obtained using the Mutual Information feature selection method with 16 miRNA features, a 60:40 training–test ratio, and the SMOTEENN balancing technique, achieving a training accuracy of 99% and a test accuracy of 96% on White patient data. However, when the model was evaluated using internal data from Black patients, the accuracy remained at 72%, suggesting that models primarily trained on White patient data may have limited generalizability across different racial groups.

Various experimental configurations were also evaluated using 17 miRNA features with different feature selection methods, training–test split ratios, and

Table VIII. *Results after using 17 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
17	Mutual Information	60/40	SMOTEENN	97%	83%	86%	96%	90%	72%
17	Mutual Information	60/40	SMOTE	83%	83%	86%	95%	90%	72%
17	Student's *t*-test	80/20	SMOTEENN	91%	82%	86%	95%	90%	72%
17	Chi-square	80/20	BorderlineSMOTE	87%	82%	88%	92%	90%	72%
17	Student's *t*-test	60/40	SMOTE	89%	81%	86%	92%	89%	72%

Acc: Accuracy; Prec: precision.

Table IX. *Results after using 18 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
18	Student's *t*-test	70/30	BorderlineSMOTE	93%	82%	86%	96%	90%	72%
18	Student's *t*-test	80/20	RandomOversampler	91%	82%	86%	95%	90%	72%
18	*T*-test	80/20	SMOTE	90%	82%	86%	95%	90%	72%
18	Chi-square	80/20	RandomOversampler	89%	82%	88%	92%	90%	72%
18	Student's *t*-test	70/30	SMOTEENN	98%	81%	85%	94%	89%	72%

Acc: Accuracy; Prec: precision.

Table X. *Results after using 19 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
19	Chi-square	80/20	SMOTE	86%	83%	88%	94%	91%	72%
19	Student's *t*-test	70/30	SMOTE	90%	82%	86%	96%	90%	72%
19	Student's *t*-test	80/20	SMOTE	90%	82%	86%	95%	90%	72%
19	Student's *t*-test	80/20	RandomOversampler	91%	82%	86%	95%	90%	72%
19	Mutual information	60/40	RandomOversampler	94%	81%	85%	94%	89%	72%

Acc: Accuracy; Prec: precision.

balancing techniques. As shown in Table VIII, the best performance was obtained using the Mutual Information feature selection method with the SMOTEENN balancing technique at a 60:40 ratio, achieving a training accuracy of 97% and a test accuracy of 83% on White patient data. Other configurations using Mutual Information with SMOTE produced comparable results, while models using Student's *t*-test and chi-square with different balancing methods such as SMOTEENN and BorderlineSMOTE resulted in slightly lower performance. Despite these variations in feature selection, balancing methods, and data split ratios, the overall accuracy remained relatively consistent across configurations. However, when evaluated using internal data from Black patients, the accuracy remained at 72%, indicating that the model trained predominantly on White patient data may have limited generalizability across different racial groups.

Various model configurations were also tested using 18 miRNA features with different feature selection methods, data split ratios, and balancing techniques. As shown in Table IX, the highest performance was achieved using the Student's *t*-test feature selection method with the BorderlineSMOTE balancing technique at a 70:30 training–test ratio, resulting in a training accuracy of 93% and a test

Table XI. *Results after using 20 miRNA features.*

Number of miRNA	Feature Selection	Ratio	Balancing	CV Test Acc	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
20	Limma	70/30	RandomOversampler	86%	96%	86%	89%	96%	92%	74%
20	Limma	70/30	SMOTE	87%	95%	85%	89%	95%	92%	74%
20	Limma	70/30	BorderlineSMOTE	88%	95%	83%	89%	92%	91%	74%
20	Limma	60/40	RandomOversampler	85%	98%	83%	87%	94%	90%	74%
20	Limma	80/20	SMOTEENN	89%	99%	82%	88%	92%	90%	74%

Acc: Accuracy; Prec: precision.

accuracy of 82% on White patient data. Other configurations using Student's *t*-test with RandomOverSampler and SMOTE at an 80:20 ratio produced similar test accuracies, while the Chi-square feature selection method combined with RandomOverSampler yielded comparable results. Meanwhile, the use of SMOTEENN slightly reduced the test accuracy despite achieving higher training accuracy. When evaluated using internal data from Black patients, the accuracy remained at 72%, indicating that the model trained primarily on White patient data may have limited generalizability across different racial groups.

Various experimental configurations were also evaluated using 19 miRNA features with different feature selection methods, training–test split ratios, and balancing techniques. As shown in Table X, the best performance was achieved using the chi-square feature selection method with the SMOTE balancing technique at an 80:20 training–test ratio, resulting in a training accuracy of 86% and a test accuracy of 83% on White patient data. Other configurations using the Student's *t*-test method with SMOTE and RandomOverSampler produced slightly lower but comparable results across both 70:30 and 80:20 ratios, while Mutual Information combined with RandomOverSampler at a 60:40 ratio resulted in the lowest test accuracy among the tested scenarios. However, when evaluated using internal data from Black patients, the accuracy remained at 72%, indicating that the model trained primarily on White patient data may have limited generalizability across different racial groups.

As shown in Table XI, although the same feature selection method, Limma, was applied, the model's performance remained relatively strong with 20 selected miRNA features. The best scenario was obtained using an 80/20 data split ratio with the SMOTEENN balancing technique, achieving the highest cross-validation accuracy of 89% and a training accuracy of 99%, while the test accuracy on White patient data reached 82%, precision of 88%, recall of 92% and F1-score of 90%. The good balance of accuracy shows strong performance especially in minority class with high F1-score. Other configurations using RandomOverSampler, SMOTE, and BorderlineSMOTE with 60/40 and 70/30 ratios produced slightly lower but comparable results. When evaluated using internal data from Black patients, the model achieved an accuracy of 74%, which is slightly higher than the results obtained in previous configurations using fewer features. The model with 20 miRNA features was also evaluated using cross-validation to further assess its stability. The cross-validation accuracy was slightly lower than the training accuracy obtained from the conventional train–test split, which may be attributed to variations across the folds during the cross-validation process. Nevertheless, this result still supports the model's overall reliability, indicating that the model maintains consistent performance when evaluated across multiple subsets of the data rather than relying on a single train–test partition.

To improve the interpretability of the XGBoost model, LIME was applied to analyse which miRNA features contributed most to individual classification decisions. Furthermore, StandardScaler was applied to standardize the feature values prior to model training, ensuring that all features were on a comparable scale

and preventing features with larger magnitudes from disproportionately influencing the model and the LIME distance-based sampling process. Figure 2 illustrates the LIME explanation for a correctly predicted prostate cancer case. The green bars represent miRNAs that support the predicted class, while the red bars indicate miRNAs that contribute toward the opposite class. Several miRNAs show strong influence on the model's decision. For example, hsa-mir-185, hsa-mir-3170, hsa-mir-7702, and hsa-mir-140 positively contributed to the predicted classification, as indicated by the green bars. In contrast, miRNAs such as hsa-mir-4716, hsa-mir-6876, hsa-mir-30b, hsa-mir-184, and hsa-mir-7109 exhibited negative contributions, represented by the red bars, suggesting their influence toward the alternative class. This explanation highlights how specific miRNA expression values contribute to the model's prediction and provides insight into the factors influencing the classification outcome.

The list of expression values for each feature demonstrates their activity in the selected sample. This patient specific explanation reveals that several highly weighted miRNAs, such as hsa-mir-185, hsa-mir-3170, hsa-mir-7702, and hsa-mir-140, contributed positively to the model's decision for early-stage classification. Such interpretability helps ensure that the model's predictions are not only accurate but also biologically meaningful, aligning with known roles of these miRNAs in cancer progression. Figure 3 shows the LIME feature ranking for 20 miRNA features that have the greatest contribution to the model's prediction. The bar chart illustrates the importance of each miRNA, where higher values indicate a stronger influence on the classification outcomes.

Discussion

This model utilizes XGBoost along with several data splitting ratios and different balancing techniques, including RandomOversampler, SMOTE, SMOTEENN, and BorderlineSMOTE. By applying various scenarios in the model-building process, a range of results were obtained, with performance improving across configurations. In addition, feature selection methods such as Student's *t*-test, mutual information, chi-square and limma, were used to identify significant miRNA features, further enhancing the potential for optimal outcomes. The results indicate that the model's performance is highly dependent on the combination of feature selection method, train-test split ratio, and balancing technique. This approach has proven effective in improving the model's accuracy in classifying cancer data more precisely.

Across all experiments, the best performance was achieved when using 20 miRNA features, with the RandomOversampler balancing technique and a 70:30 train-test ratio, resulting in an accuracy of 86%, precision of 89%, recall of 96%, and F1-score of 92%. The results with high F1-score suggest the model is particularly strong in identifying minority class. While tested using cross validation, 20 miRNA features also has the higher accuracy but with different balancing technique and ratio. The balancing technique that was being used is SMOTEENN with ratio of 80:20, resulting in accuracy of 89%. The high metric values indicate that the model not only excels at identifying the majority class but is also highly sensitive to the minority class. These findings highlight the critical role of balancing techniques and feature selection in the development of genomic-based classification models.

However, when the same model was tested on data from Black patients, the test accuracy dropped to 74%. This decline indicates that a model trained predominantly on data from White patients has limitations in generalizing its performance across different racial groups. These results emphasize that race is a relevant factor in the performance of genomic-based classification models and underscore the importance of considering population diversity in the training process to build fairer and more inclusive predictive systems.

Feature selection plays a crucial role in determining the final performance of a classification model. In this study, 4 feature selection scenarios were implemented. Scenario 1 used a bioinformatics approach with limma, applying the criteria of p-Value <0.01 to select significant

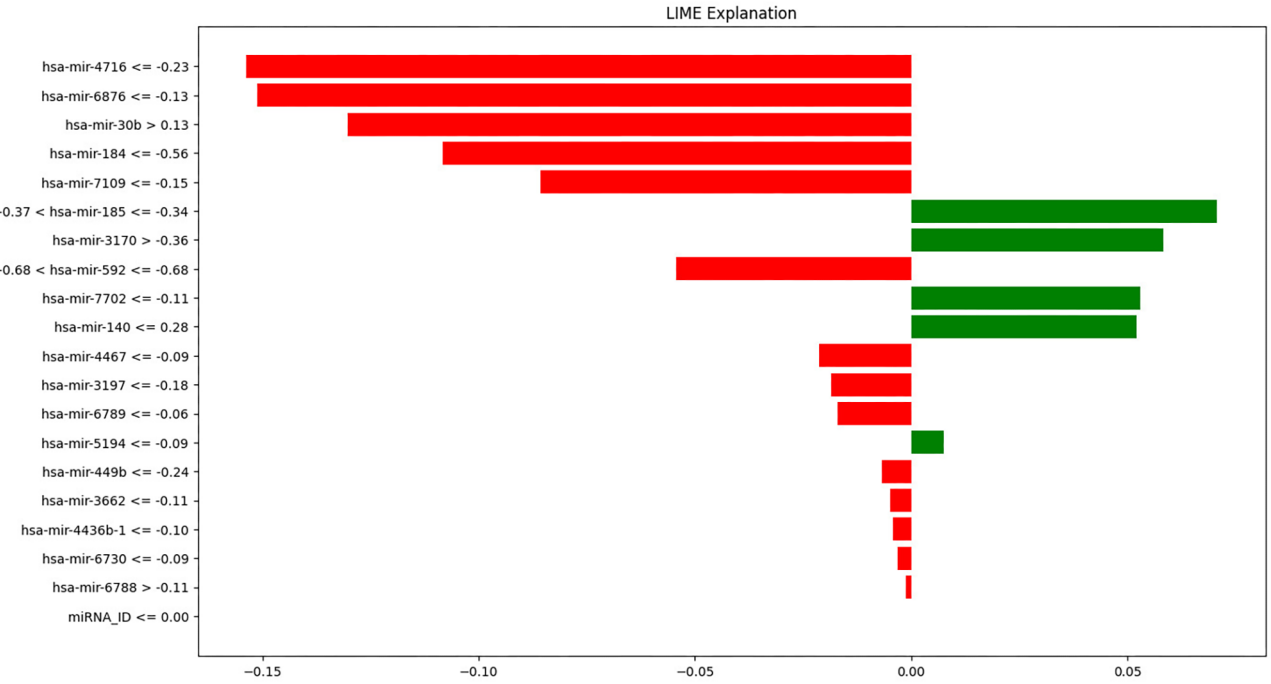

Figure 2. *Local interpretable model-agnostic (LIME) explanations illustrating the contribution of selected miRNA features to early and late prostate cancer classification. Green bars represent miRNAs that support the predicted class, while the red bars indicate miRNAs that contribute toward the opposite class.*

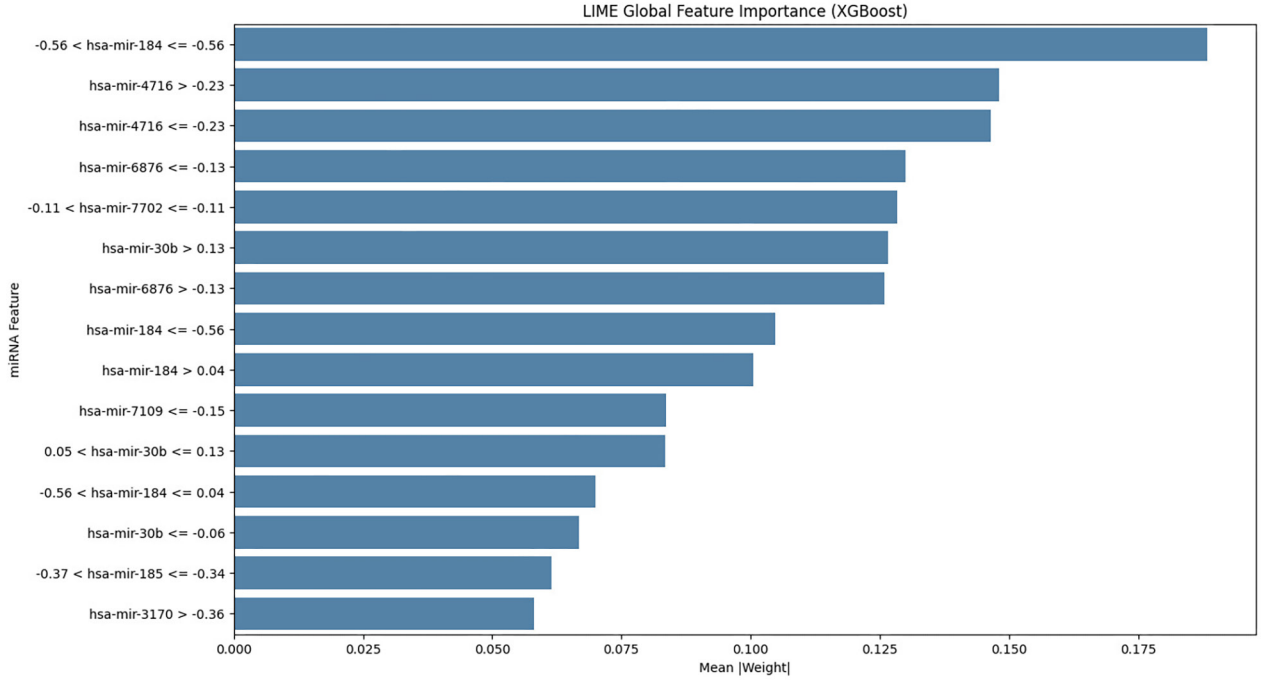

Figure 3. *Ranking of miRNA features based on the mean absolute local interpretable model-agnostic explanations (LIME) weight in the XGBoost model for prostate cancer stage classification.*

miRNA features, resulting in the highest accuracy of 89%. Scenarios 2 to 4 applied a statistical feature selection such as Student's *t*-test, mutual information and chi-square, with variations in the maximum number of selected miRNA features, *i.e.* 15, 16, 17, 18, and 19 miRNAs. All three scenarios achieved the accuracy between 83% to 85%, demonstrating the consistent performance of the statistical feature selection approach.

Based on the models developed using various scenarios, the set of 20 miRNA features demonstrated the best overall performance. Therefore, the implementation for obtaining these 20 miRNA features can be carried out using next-generation sequencing (NGS) technology, specifically with the Illumina HiSeq 2000 platform, to align with the sequencing approach used in the existing training data. During the implementation process, the sequencing results yielded 1,881 normalized miRNA expression profiles in the form of reads per million (RPM). The identified miRNAs were then matched against the selected 20 miRNA features to ensure feature consistency, enabling effective detection of prostate cancer.

Conclusion

This study developed a prostate cancer detection model using the XGBoost algorithm, taking into account miRNA data and racial factors. Feature selection methods such as Student's *t*-test, mutual information, chi-square and limma were employed to identify significant miRNA features for model development. The results show that with proper feature selection, the model achieved high accuracy of up to 89%, with a precision of 89%, recall of 96% and F1-score of 92%, demonstrating that feature selection plays a crucial role in improving predictive performance. Additionally, when tested on data from Black patients, the model achieved an accuracy of 74%. This difference indicates a possible variation in miRNA expression patterns across races, which should be considered when developing more inclusive and representative classification models.

Conflicts of Interest

The Authors declare no conflicts of interest.

Authors' Contributions

Conceptualization: DA, DAK, AKW, JS; Data curation: JS, AM, VK; Formal analysis: JS; Funding acquisition: DA; Investigation: JS; Methodology: DA, DAK, AKW, JS, EBW, RN; Project administration: JS, AM, VK; Resources: JS, AM, VK; Supervision: DA, DAK, AKW, EBW, RN; Validation: DA, DAK, AKW, EBW, RN; Visualization: JS, AM, VK; Writing – original draft: JS, AM, VK; Writing – review & editing: JS, AM, VK.

Acknowledgements

The Authors wish to express their sincere gratitude to Universitas Multimedia Nusantara (UMN) for their continuous support and encouragement, which greatly facilitated the successful completion of this research.

Funding

This research was funded by the Indonesian Government under Kemdiktisaintek penelitian fundemental reguler schema with contract number 0993/LL3/AL.04/2025.

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 Wang L, Lu B, He M, Wang Y, Wang Z, Du L: Prostate cancer incidence and mortality: Global status and temporal trends in 89 countries from 2000 to 2019. Front Public Health 10: 811044, 2022. DOI: 10.3389/fpubh.2022.811044

2 Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71(3): 209-249, 2021. DOI: 10.3322/caac.21660

3 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

4 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

5 Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13: 8-17, 2014. DOI: 10.1016/j.csbj.2014.11.005

6 Ma B, Meng F, Yan G, Yan H, Chai B, Song F: Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med 121: 103761, 2020. DOI: 10.1016/j.compbiomed.2020.103761

7 Arif Ali Z, Abduljabbar ZH, Tahir HA, Bibo Sallow A, Almufti SM: eXtreme gradient boosting algorithm with machine learning: a review. Acad J Nawroz Univ 12(2): 320-334, 2023. DOI: 10.25007/ajnu.v12n2a1612

8 Liew XY, Hameed N, Clos J: An investigation of XGBoost-based algorithm for breast cancer classification. Mach Learn Appl 6: 100154, 2021. DOI: 10.1016/j.mlwa.2021.100154

9 Tarwidi D, Pudjaprasetya SR, Adytia D, Apri M: An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. MethodsX 10: 102119, 2023. DOI: 10.1016/j.mex.2023.102119

10 Guan X, Du Y, Ma R, Teng N, Ou S, Zhao H, Li X: Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. BMC Med Inform Decis Mak 23(1): 107, 2023. DOI: 10.1186/s12911-023-02171-x

11 Loa J, Wiratama J, Halim FA: Optimizing inventory management in retail companies through sales prediction using XGBoost. 2025 4th International Conference on Electronics Representation and Algorithm (ICERA): 1-6, 2025. DOI: 10.1109/ICERA66156.2025.11087347

12 Juliandri R, Johan ME, Wiratama J, Sanjaya SA: Adverse Media Classification: A New Era of Risk Management with XGBoost and Gradient Boosting Algorithms. 2024 5th International Conference on Big Data Analytics and Practices (IBDAP): 18-21, 2024. DOI: 10.1109/IBDAP62940.2024.10689708

13 Ogunleye A, Wang QG: XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform 17(6): 2131-2140, 2020. DOI: 10.1109/TCBB.2019.2911071

14 Agustriawan D, Kurniawan V, Overbeek MV, Widjaja M, Mulia A, Syechlo J, Ahmad MI, Daryanto B, Seputra KP, Susilo H, Negara EP, Effendi RA, Sathipati SY: Robust logistic regression-based diagnosis method of prostate cancer using optimized feature selection on race specific gene-expression datasets. Cancer Diagn Progn 5(6): 720-734, 2025. DOI: 10.21873/cdp.10487

15 Lu TX, Rothenberg ME: MicroRNA. J Allergy Clin Immunol 141(4): 1202-1207, 2018. DOI: 10.1016/j.jaci.2017.08.034

16 Reddy KB: MicroRNA (miRNA) in cancer. Cancer Cell Int 15: 38, 2015. DOI: 10.1186/s12935-015-0185-1

17 Smolarz B, Durczyński A, Romanowicz H, Szyłło K, Hogendorf P: miRNAs in cancer (review of literature). Int J Mol Sci 23(5): 2805, 2022. DOI: 10.3390/ijms23052805

18 Wu HH, Leng S, Sergi C, Leng R: How microRNAs command the battle against cancer. Int J Mol Sci 25(11): 5865, 2024. DOI: 10.3390/ijms25115865

19 Champeris Tsaniras S, Delinasios GJ, Petropoulos M, Panagopoulos A, Anagnostopoulos AK, Villiou M, Vlachakis D, Bravou V, Stathopoulos GT, Taraviras S: DNA replication inhibitor geminin and retinoic acid signaling participate in complex interactions associated with pluripotency. Cancer Genomics Proteomics 16(6): 593-601, 2019. DOI: 10.21873/cgp.20162

20 Rishabh K, Khadilkar S, Kumar A, Kalra I, Kumar AP, Kunnumakkara AB: MicroRNAs as modulators of oral tumorigenesis-a focused review. Int J Mol Sci 22(5): 2561, 2021. DOI: 10.3390/ijms22052561

21 Komatsu S, Kitai H, Suzuki HI: Network regulation of microRNA biogenesis and target interaction. Cells 12(2): 306, 2023. DOI: 10.3390/cells12020306

22 O'Brien J, Hayder H, Zayed Y, Peng C: Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne) 9: 402, 2018. DOI: 10.3389/fendo.2018.00402

23 Haseltine WA, Patarca R: The RNA revolution in the central molecular biology dogma evolution. Int J Mol Sci 25(23): 12695, 2024. DOI: 10.3390/ijms252312695

24 Ali Syeda Z, Langden SSS, Munkhzul C, Lee M, Song SJ: Regulatory mechanism of microRNA expression in cancer. Int J Mol Sci 21(5): 1723, 2020. DOI: 10.3390/ijms21051723

25 Kanli A, Sunnetci-Akkoyunlu D, Kulcu-Sarikaya N, Ugurtaş C, Akpinar G, Kasap M: Potential common molecular mechanisms between periodontitis and prostate cancer: a network analysis of differentially expressed miRNAs. In Vivo 39(2): 795-809, 2025. DOI: 10.21873/invivo.13863

26 Lowder D, Rizwan K, McColl C, Paparella A, Ittmann M, Mitsiades N, Kaochar S: Racial disparities in prostate cancer: A complex interplay between socioeconomic inequities and

genomics. Cancer Lett 531: 71-82, 2022. DOI: 10.1016/j.canlet.2022.01.028
27 Bigler SA, Pound CR, Zhou X: A retrospective study on pathologic features and racial disparities in prostate cancer. Prostate Cancer 2011: 239460, 2011. DOI: 10.1155/2011/239460
28 Ning Z, Yu S, Zhao Y, Sun X, Wu H, Yu X: Identification of miRNA-mediated subpathways as prostate cancer biomarkers based on topological inference in a machine learning process using integrated gene and miRNA expression data. Front Genet 12: 656526, 2021. DOI: 10.3389/fgene.2021.656526
29 Bergez-Hernández F, Arámbula-Meraz E, Alvarez-Arrazola M, Irigoyen-Arredondo M, Luque-Ortega F, Martínez-Camberos A, Cedano-Prieto D, Contreras-Gutiérrez J, Martínez-Valenzuela C, García-Magallanes N: Expression analysis of miRNAs and their potential role as biomarkers for prostate cancer detection. Am J Mens Health 16(5): 15579883221120989, 2022. DOI: 10.1177/15579883221120989
30 Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, Zhu J, Haussler D: Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 38(6): 675-678, 2020. DOI: 10.1038/s41587-020-0546-8
31 AJCC Cancer Staging Manual. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A (eds.). New York, NY, USA, Springer, 2010.
32 Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507-2517, 2007. DOI: 10.1093/bioinformatics/btm344
33 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7): e47, 2015. DOI: 10.1093/nar/gkv007
34 Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR: Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. J Big Data 11(1), 2024. DOI: 10.1186/s40537-023-00857-7
35 Husain G, Nasef D, Jose R, Mayer J, Bekbolatova M, Devine T, Toma M: SMOTE vs. SMOTEENN: a study on the performance of resampling algorithms for addressing class imbalance in regression models. Algorithms 18(1): 37, 2025. DOI: 10.3390/a18010037
36 Sun Y, Que H, Cai Q, Zhao J, Li J, Kong Z, Wang S: Borderline SMOTE algorithm and feature selection-based network anomalies detection strategy. Energies 15(13): 4751, 2022. DOI: 10.3390/en15134751
37 Santos MS, Soares JP, Abreu PH, Araujo H, Santos J: Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Comput Intell Mag 13(4): 59-76, 2018. DOI: 10.1109/MCI.2018.2866730