

# Prostate Cancer Stage Classification Using the XGBoost Model on miRNA Data

David Agustriawan<sup>1\*</sup>, Jheno Syechlo<sup>1</sup>, Marlinda Vasty Overbeek<sup>1</sup>, Vincent Kurniawan<sup>1</sup>, Adithama Mulia<sup>1</sup>, Moeljono Widjaja<sup>1</sup>, Muhammad Imran Ahmad<sup>2</sup>, Srinivasulu Yerukala Sathipati<sup>3</sup>

<sup>1</sup>Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Indonesia.

<sup>2</sup>Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia.

<sup>3</sup>Marshfield Clinic Research Institute, Marshfield Clinic Research Institute, Marshfield Wisconsin 54449

\*Corresponding Author. Email: david.agustriawan@umn.ac.id

## Abstract

This study aims to develop a prostate cancer stage classification model using the XGBoost algorithm on miRNA expression data, with a focus on race-based analysis. The choice of this topic is based on the high mortality rate from prostate cancer and diagnostic disparities between races, particularly between White and Black populations. The data used were obtained from the GDC TCGA XenaBrowser, comprising miRNA expression and patient clinical data. This study applied several feature selection methods such as Lasso + RFE, edgeR, and ROC, as well as data balancing techniques including RandomOversampler, SMOTE, SMOTEEN, and BorderlineSMOTE. The results show that the XGBoost model achieved an accuracy of up to 99% on data from White patients. However, when tested on data from Black patients, accuracy decreased to 84–89%, indicating limitations in cross-race performance. The main challenge in this study was the limited amount of data and the approach used to compare performance across races to identify the impact of race on cancer stage classification. This study is expected to serve as an initial foundation for developing more inclusive and equitable cancer classification models.

**Keywords:** prostate cancer, miRNA, XGBoost, feature selection, racial disparity

## 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]. Prostate cancer is the most frequently diagnosed cancer in 112 countries, and the leading cause of cancer death in 48 countries. Currently, machine learning techniques have been used in research to further develop the diagnosis of prostate cancer [2], [3]. Machine learning techniques can discover patterns from complex datasets and effectively predict the outcome of prostate cancer [4].

XGBoost, also known as eXtreme Gradient Boosting, is one of many machine learning algorithms that has more advanced implementation than gradient boosting [5], [6]. The algorithm is called "Extreme" because it uses regularization to prevent overfitting [7]. 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 [8], [9]. 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 demonstrated the strong performance of XGBoost in liver disease, achieving high accuracy and sensitivity [10]. This makes XGBoost become a robust tool for detection of prostate cancer. While XGBoost algorithms can enhance the accuracy of prostate cancer, the outcome is also affected by other key factors. One of the key factors is race disparity in prostate cancer.

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 [11], [12]. miRNAs control various biological processes such as cell division, cell differentiation, angiogenesis, migration, apoptosis, and oncogenesis [13], [14]. 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 [15]. Other than that, each miRNA can have multiple targeted genes [16], [17]. This broad targeting allows miRNAs to regulate complex biological pathways. miRNAs are also related to the concept of the central dogma. The concept of the central dogma describes the stages by which DNA is broken down and processed into proteins [18]. 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 [19].

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 [20]. Furthermore, Black men are more likely to be diagnosed at a more advanced stage of the disease [21]. 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.

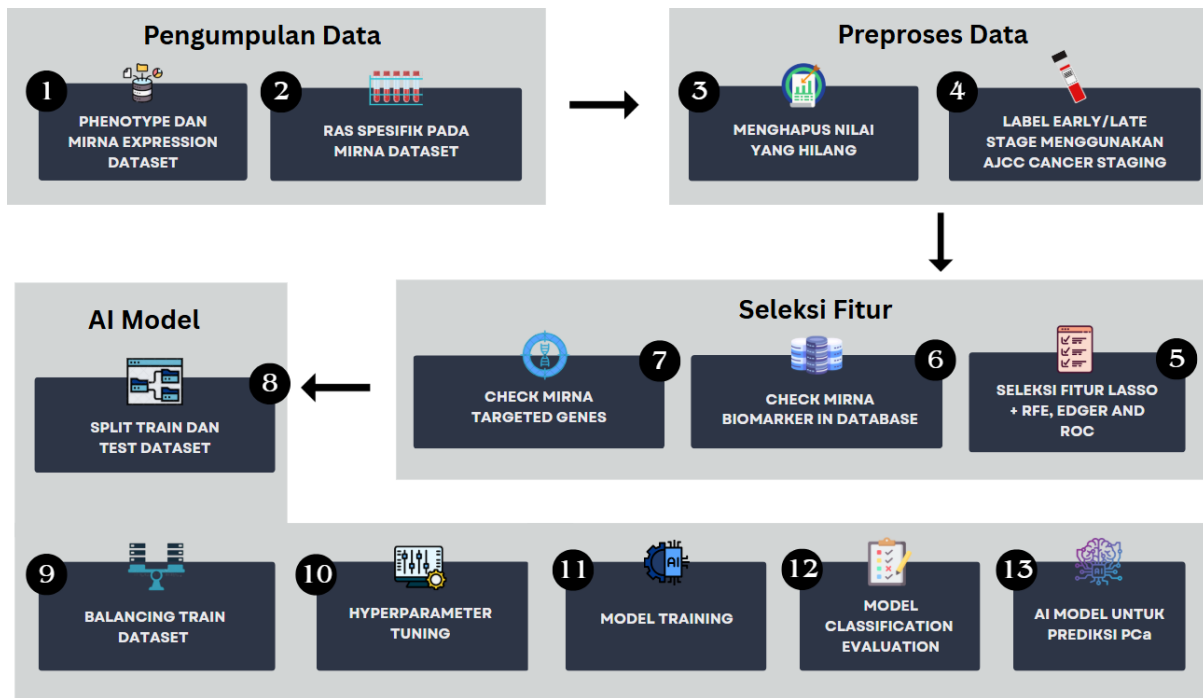
In previous studies, there has not been any research that used miRNA data with the XGBoost algorithm. A study conducted by Kalaiyarasi M. et al. used a miRNA dataset on prostate cancer. This study performed feature selection from a website to identify the most significant miRNAs, resulting in 209 miRNAs. The study achieved an accuracy of 92% and an AUC of 95% using the SVM algorithm [22]. Other study conducted by Fernando et al. used 4 miRNAs specifically without feature selection to get the accuracy for biomarker in prostate cancer. The study achieved precision of 0.763 and accuracy of 0.762 by using logistic regression [23].

Unlike previous research, the current study uses data obtained from the GDC TCGA XenaBrowser and applies the XGBoost algorithm, which has not yet been used in prostate cancer staging detection. Furthermore, this study employs different feature selection methods to identify significant miRNA features, such as EdgeR, Lasso + RFE, and ROC, and also performs AUC Score validation. The study uses various balancing techniques to achieve balanced data. The data used focuses on patients of the white race, and testing will be conducted on black race patients to observe the impact of racial differences. By using a combination of different algorithms and feature selection methods, this research aims to build a more accurate classification model for prostate cancer staging, allowing patients to detect prostate cancer at an earlier stage.

## **Methodology**

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.

All steps in this study were conducted by python version 3.13.2 in visual studio software. The devices used for this study included a Windows 11 OS, dengan 16GB RAM, processor 12th Gen Intel® Core™ i5-12500H (16 CPUs), 2.5Hz dan NVIDIA GeForce RTX 3050 4GB.



*Figure 1 Research Pipeline for Prostate Cancer*

## 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 [24]. 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  $\log_2(\text{RPM} + 1)$  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. The

rules used for label grouping are detailed in Table 1 [25]. 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 labeling process.

Table 1. Prostate Cancer Staging Criteria

Group	T	N	M	PSA	Gleason
I	T1a–c	N0	M0	PSA < 10	Gleason $\leq 6$
	T2a	N0	M0	PSA < 10	Gleason $\leq 6$
	T1–2a	N0	M0	PSA X	Gleason X
IIA	T1a–c	N0	M0	PSA < 20	Gleason 7
	T1a–c	N0	M0	PSA $\geq 10$ < 20	Gleason $\leq 6$
	T2a	N0	M0	PSA $\geq 10$ < 20	Gleason $\leq 6$
	T2a	N0	M0	PSA < 20	Gleason 7
	T2b	N0	M0	PSA < 20	Gleason $\leq 7$
	T2b	N0	M0	PSA X	Gleason X
IIB	T2c	N0	M0	Any PSA	Any Gleason
	T1–2	N0	M0	PSA $\geq 20$	Any Gleason
	T1–2	N0	M0	Any PSA	Gleason $\geq 8$
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

## Feature Selection

After obtaining data labeled as early and late stage, feature selection was conducted to identify the most significant and cancer-related miRNA features. RFE and Lasso were used to identify significant features by specifying the desired number of features in the RFE method. In addition, a bioinformatics-based feature selection technique, EdgeR, was also used to identify significant miRNA features. Feature selection using EdgeR focused on the logFC values and p-values. The feature selection methods used to identify significant miRNA features are summarized in Table 2. Scikit-learn version 1.5.1 was used for implementing RFE and Lasso, while EdgeR was run using version 4.0.16. ROC analysis was also employed from the beginning to select significant miRNA features [26], [27]. The identified miRNA expressions will be verified using the websites miRCancer and dbDEMC (Database of Differentially Expressed miRNAs in Human Cancers) [28], [29]. Several miRNAs that show a strong correlation with prostate cancer will be analysed to identify their target genes using miRDB [30]. Through this approach, the study aims to determine whether the genes targeted by these miRNAs are also associated with cancer development or progression.

Table 2. Feature Selection Method

No.	Feature Selection Method	Criteria
1	EdgeR	PValue < 0.01 and logFC > 0.4
2	Lasso + RFE	Best result from 15, 18, and 19 miRNA features
3	ROC	ROC > 0.6

## 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. The package used for data balancing was imbalanced-learn version 0.12.3.

The model involved several scenarios focusing on feature selection methods, including RFE, Lasso, ROC, and EdgeR. 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. After completing all scenarios, validation was conducted using the AUC score. The scenarios used in the model development process are presented in Table 3.

Table 3. Modelling Scenario Using XGBoost

<i>Feature Selection</i>	<i>Splitting Ratio</i>	<i>Balancing Technique</i>
EdgeR criteria PValue < 0.01 and logFC > 0.4	80:20 70:30 60:40	<i>Random Oversampler</i> <i>SMOTEEN</i> <i>KMeans SMOTE</i> <i>Borderline SMOTE</i>
Lasso + RFE with the best result from 15, 18 dan 19 feature	80:20 70:30 60:40	<i>Random Oversampler</i> <i>SMOTEEN</i> <i>KMeans SMOTE</i> <i>Borderline SMOTE</i>
ROC criteria > 0.6	80:20 70:30 60:40	<i>Random Oversampler</i> <i>SMOTEEN</i> <i>KMeans SMOTE</i> <i>Borderline SMOTE</i>

The resulting model is evaluated using a classification report and confusion matrix, which include 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 model development process is illustrated in Figure 3. The finalized model is then tested using a different dataset specifically, data from Black patients to assess the significance of racial differences.

## Result

### 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 4. 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.

Table 4. Dataset miRNA

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

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 modeling process.

## Feature Selection

Feature selection was performed using two different approaches: edgeR and Lasso + RFE. The edgeR method resulted in a total of 3 selected miRNAs. The criteria used in this scenario were  $p\text{-value} < 0.01$  and  $\log\text{FC} > 0.4$ . In addition, feature selection using Lasso and RFE yielded 15, 18, and 19 miRNAs. With Lasso and RFE, the number of selected features can be determined based on the desired maximum. The selected miRNA features showed significant expression results. To further improve performance, ROC analysis was also applied with a threshold  $> 0.6$ , resulting in 7 miRNAs. The results of feature selection are presented in Table 5. The selected features represent the most optimal set for model construction.

Table 5. Feature Selection Results

No.	Feature Selection Method	Criteria	miRNA Feature
1	EdgeR	$P\text{Value} < 0.01$ and $\log\text{FC} > 0.4$	3 miRNA
2	Lasso + RFE	Best result from 15 miRNA feature	15 miRNA



3	Lasso + RFE	Best result from 18 miRNA feature	18 miRNA
4	Lasso + RFE	Best result from 159miRNA feature	19 miRNA
5	ROC	AUC > 0.6	7 miRNA

The miRNA expressions obtained from various feature selection scenarios will be verified using several websites such as miRCancer and dbDEMC. These websites will be used to check whether the miRNA biomarkers are associated with prostate cancer. The biomarkers can be seen in Table 6.

Tabel 6. Biomarker miRNA pada Database

No.	miRNA ID	<i>miRCancer</i>		<i>dbDEMC</i>	
		Biomarker	Prostate Cancer	Biomarker	Prostate Cancer
1.	hsa-mir-21	V	V	V	V
2.	hsa-mir-302a	V	V	V	V
3.	hsa-mir-3155a	X	X	V	V
4.	hsa-mir-3662	V	X	V	V
5.	hsa-mir-370	V	X	V	V
6.	hsa-mir-4436b-1	X	X	X	X
7.	hsa-mir-4532	X	X	V	V
8.	hsa-mir-4673	X	X	V	V
9.	hsa-mir-4771-2	X	X	X	X
10.	hsa-mir-4795	X	X	V	V
11.	hsa-mir-490	V	X	V	V
12.	hsa-mir-498	V	X	V	V
13.	hsa-mir-555	X	X	V	V
14.	hsa-mir-631	X	X	V	V
15.	hsa-mir-6504	X	X	V	V
16.	hsa-mir-6761	X	X	V	V
17.	hsa-mir-6785	X	X	V	V
18.	hsa-mir-6876	X	X	V	V
19.	hsa-mir-7641-1	X	X	X	X

Table 3 illustrates that the biomarkers identified through various feature selection scenarios exhibit correlations with prostate cancer when evaluated using three different databases such as miRCancer, and dbDEMC. The verification process involved assessing each miRNA biomarker for general cancer association as well as its specific relevance to prostate cancer. While some miRNAs were highly associated with cancer, others demonstrated a direct and strong link to prostate cancer. hsa-mir-21 showed a strong and correlation across all databases, supporting its significance as a potential biomarker for prostate cancer. Conversely, several biomarkers were found to be associated with other cancer types but not with prostate cancer. In certain cases, some miRNAs were not detected in specific databases, likely due to the absence of those entries in the database repositories. For example, hsa-mir-3155a was only found in the dbDEMC database. Among the platforms used, dbDEMC demonstrated the most comprehensive corellation, detecting a greater number of relevant biomarkers during the validation process. This suggests that dbDEMC may offer a more extensive and up to date resource for miRNA-cancer associations.

Every miRNA in human can have targeted gene to regulate gene expression. For example, hsa-miR-21 targets the SCAI gene, which acts as a suppressor of cancer cell invasion. When miR-21 reduces the expression of SCAI, it can make it easier for cancer cells to spread. Another example is hsa-miR-302a, which targets the CASC1 gene which linked to a higher risk of developing certain types of cancer. These cases show that miRNAs can control genes that are important in cancer. By targeting multiple genes, miRNAs can help suppress or sometimes support cancer development.

## 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.

Table 7. Results for using 3 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
3	EdgeR	60/40	Random Oversampler	100%	98%	99%	98%	100%	99%	88%

3	EdgeR	80/20	Random Oversampler	100%	98%	99%	98%	100%	99%	88%
3	EdgeR	70/30	Random Oversampler	100%	98%	98%	98%	100%	99%	88%
3	EdgeR	60/40	Borderlin eSMOTE	93%	91%	92%	93%	98%	95%	86%
3	EdgeR	60/40	SMOTE	92%	83%	92%	92%	100%	96%	88%

As shown in Table 7, the use of RandomOversampler yielded the best results when applied with 3 miRNA features. A training-to-test data ratio of 60/40 produced the highest accuracy at 99%. In addition, the results demonstrated consistent performance despite changes in the data split ratios (60/40, 70/30, and 80/20), with accuracy remaining relatively high. This indicates that the choice of data split ratio does not significantly affect model accuracy, as long as the number of features and balancing technique remain the same. Furthermore, an AUC score of 100% highlights the model's ability to perfectly distinguish between classes. However, when the model was tested using internal data from Black patients, the resulting accuracy dropped to 88%, indicating that a model trained primarily on data from White patients has limited generalizability to other racial groups.

Table 8. Results for using 15 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
15	Lasso + RFE	60/40	<i>RandomOversampler</i>	99%	98%	97%	97%	100%	98%	89%
15	Lasso + RFE	70/30	<i>RandomOversampler</i>	98%	98%	97%	97%	100%	98%	89%
15	Lasso + RFE	60/40	<i>SMOTE</i>	98%	99%	97%	97%	100%	98%	84%
15	Lasso +	80/20	<i>SMOTE</i>	98%	98%	96%	95%	100%	98%	84%

	RFE									
15	Lasso + RFE	80/20	<i>RandomOversampler</i>	98%	98%	96%	95%	100%	98%	89%

As shown in Table 8, the use of 15 miRNA features selected through the Lasso and RFE methods resulted in a relatively high accuracy performance, reaching up to 97%. The best results were obtained with a 60/40 training-test ratio using the RandomOversampler balancing technique. The results indicate that this feature selection combination consistently delivers strong model performance, even with variations in the data split ratios and balancing techniques applied. The high AUC score of up to 99% further confirms the model's excellent classification ability in distinguishing between classes. However, when tested on data from Black patients, the model's accuracy dropped to between 84% and 89%, highlighting its limitations in generalizing to different racial groups.

Table 9. Results for using 16 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
16	Lasso + RFE	60/40	<i>RandomOversampler</i>	99%	98%	97%	96%	90%	93%	89%
16	Lasso + RFE	80/20	<i>RandomOversampler</i>	98%	98%	96%	97%	85%	90%	89%
16	Lasso + RFE	70/30	<i>SMOTE</i>	97%	94%	97%	95%	93%	95%	89%
16	Lasso + RFE	80/20	<i>SMOTE</i>	97%	94%	97%	98%	90%	93%	89%
16	Lasso + RFE	70/30	<i>RandomOversampler</i>	98%	98%	97%	96%	90%	93%	89%

As shown in table 9, the results demonstrates that the use of 16 miRNA features selected through the combination of Lasso and Recursive Feature Elimination (RFE) consistently yields excellent model performance, with AUC scores ranging from 97% to 99% across all configurations. The configuration employing a 60/40 training-to-testing data ratio and the RandomOversampler technique achieved the highest AUC score of 99%, along with a test accuracy of 97%, precision of 96%, recall of 90%, and F1-score of 93%. These results indicate that the model is capable of effectively and fairly detecting the target class within the test data. Furthermore, the model maintained consistent accuracy at 89% when applied to data from Black patients across all configurations, highlighting its robustness and potential applicability in diverse demographic settings. Overall, the combination of Lasso + RFE feature selection and the RandomOversampler balancing method produced the best classification performance, particularly under the 60/40 and 70/30 training-testing splits. These findings underscore the critical role of data ratio configuration and balancing techniques in enhancing the predictive capability of the classification model.

Table 10. Results for using 17 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
17	Lasso + RFE	70/30	<i>RandomOversampler</i>	99%	98%	99%	99%	96%	98%	91%
17	Lasso + RFE	60/40	<i>RandomOversampler</i>	99%	98%	99%	99%	95%	97%	91%
17	Lasso + RFE	70/30	<i>RandomOversampler</i>	99%	98%	99%	99%	95%	97%	91%
17	Lasso + RFE	60/40	<i>BorderlineSMOTE</i>	98%	94%	95%	94%	85%	89%	91%
17	Lasso + RFE	60/40	<i>SMOTE</i>	93%	93%	92%	90%	76%	72%	91%

As shown in table 10, the results highlights the strong classification performance achieved using 17 miRNA features selected through Lasso and Recursive Feature Elimination (RFE),

particularly when employing the RandomOversampler technique. Under both 70/30 and 60/40 train-test splits, the model achieved an AUC score of 99%, test accuracy of up to 99%, and a high F1-score of 98%. These results demonstrate the model's excellent precision and consistency in classifying data from White patients. In contrast, the use of the SMOTE balancing technique resulted in a noticeable decline in performance, with the AUC score dropping to 93% and the F1-score falling to 72%. Interestingly, the test accuracy on data from Black patients remained consistent at 91% across all configurations, indicating the model's robustness and potential for generalization across different racial groups.

Table 11. Results for using 18 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
18	Lasso + RFE	80/20	<i>RandomOversampler</i>	99%	99%	99%	98%	100%	99%	89%
18	Lasso + RFE	70/30	<i>RandomOversampler</i>	99%	99%	99%	99%	100%	99%	89%
18	Lasso + RFE	70/30	<i>SMOTE</i>	100%	98%	99%	99%	100%	99%	89%
18	Lasso + RFE	70/30	<i>BorderlineSMOTE</i>	99%	97%	97%	98%	99%	98%	89%
18	Lasso + RFE	70/30	<i>SMOTEEN</i>	76%	95%	88%	88%	99%	93%	93%

As shown in Table 11, the use of 18 miRNA features selected through Lasso and RFE achieved the highest test accuracy of 99%, with a precision of 98%, recall of 100%, and an F1-score of 99%. This best-performing combination was obtained using a 70/30 training-test ratio with the RandomOversampler balancing technique, indicating that the model is not only accurate but also benefits from a well-balanced dataset during training. The AUC score, reaching between 99% and 100%, further confirms the model's strong classification capability. On the other hand, the SMOTEEN technique yielded relatively lower results, with a test accuracy of only 88% and an AUC score of 76%. This suggests that SMOTEEN may be less effective in handling data distribution for this configuration compared to other balancing techniques such as

RandomOversampler or SMOTE. Additionally, when tested on data from Black patients, the model's accuracy ranged from 89% to 93%, showing an improvement compared to testing with fewer selected features.

Table 12. Results for using 19 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	CV Test Acc	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
19	Lasso + RFE	60/40	<i>RandomOversampler</i>	99%	86%	99%	99%	99%	100%	100%	89%
19	Lasso + RFE	80/20	<i>RandomOversampler</i>	99%	88%	99%	99%	98%	100%	99%	89%
19	Lasso + RFE	70/30	<i>SMOTE</i>	100%	82%	98%	99%	99%	100%	99%	89%
19	Lasso + RFE	60/40	<i>SMOTE</i>	100%	85%	98%	99%	99%	100%	100%	89%
19	Lasso + RFE	60/40	<i>BorderlineSMOTE</i>	98%	88%	97%	97%	97%	99%	98%	89%

As shown in Table 12, although the same feature selection methods Lasso and RFE were used, the model's performance remained exceptionally high with 19 selected miRNA features. The best scenario was achieved with a 60/40 data split ratio and the RandomOversampler balancing technique, resulting in a test accuracy of 99%, along with a precision of 99%, recall of 100%, and an F1-score of 100%. These metric values indicate that the model is capable of classifying the data with excellent accuracy. Supported by an AUC score of 99%, the model demonstrates outstanding discriminative ability. Compared to other balancing techniques such as SMOTE or BorderlineSMOTE, RandomOversampler continued to yield superior results in model development using 19 features. When tested on data from Black patients, the model maintained an accuracy of 89% across all configurations, indicating better performance consistency across racial groups compared to models using fewer features. The model using 19 miRNA features were tested using cross validation since it has the best accuracy among all model. The accuracy for 5-fold cross validation lower than the normal splitting due to increase in variation across fold in cross validation. However, this does not contradict the high performance seen in the

main evaluation—it instead reinforces the model's overall reliability and robustness by showing that it performs well across multiple subsets of data, not just a single train-test split.

Table 13. Results for using 7 miRNA features

Number of miRNA	Feature Selection	Ratio	Balancing	AUC Score	Train Acc	Test Acc (White)	Test Prec	Test Recall	Test F1 Score	Test Acc (Black)
7	ROC	80/20	<i>SMOTE</i>	100%	97%	99%	98%	100%	99%	84%
7	ROC	60/40	<i>SMOTE</i>	99%	97%	97%	97%	100%	98%	84%
7	ROC	70/30	<i>SMOTE</i>	99%	97%	97%	97%	100%	98%	84%
7	ROC	80/20	<i>BorderlineSMOTE</i>	97%	94%	93%	94%	98%	96%	86%
7	ROC	60/40	<i>BorderlineSMOTE</i>	97%	94%	93%	93%	99%	96%	86%

As shown in Table 13, the use of ROC-based feature selection with 7 miRNA features produced excellent performance, especially when combined with SMOTE and an 80/20 data split ratio, achieving a test accuracy of 99%. Furthermore, the high precision, recall, and F1-score values indicate that the model is not only accurate but also stable in detecting both classes evenly. Despite the relatively small number of features, the model maintained strong performance across various scenarios, particularly when using SMOTE as the balancing technique. This suggests that feature selection using ROC remains effective even with low-dimensional feature sets. Supported by an AUC score of up to 100%, the model demonstrated highly optimal classification capability. However, when tested on data from Black patients, the accuracy dropped to 84%, indicating a performance gap between racial groups. This highlights that while ROC-based feature selection is effective for the main dataset, models with fewer features may struggle to generalize well to data from different racial backgrounds.

## Discussion

This model utilizes XGBoost along with several data splitting ratios and different balancing techniques, including RandomOversampler, SMOTE, SMOTEEN, 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 Lasso + RFE and edgeR 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 19 miRNA features, with the RandomOversampler balancing technique and a 60/40 train-test ratio, resulting in an accuracy of 99%, precision of 99%, recall of 100%, and F1-score of 100%. This model demonstrates that RandomOversampler consistently handles imbalanced samples effectively in the context of genetic data. 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 89%. 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, seven feature selection scenarios were implemented. Scenario 1 used a statistical approach with edgeR, applying the criteria of  $p\text{-value} < 0.01$  and  $\log\text{FC} > 0.4$  to select significant miRNA features, resulting in the highest accuracy of 99%. Scenarios 2 to 4 applied a combination of Lasso and RFE, with variations in the maximum number of selected miRNA features 15, 16, 17, 18, and 19 miRNAs, respectively. All three scenarios achieved the same accuracy of 99%, demonstrating the consistent performance of the Lasso + RFE approach. Scenario 5 employed the ROC Curve method and selected 7 miRNAs with the highest AUC values, which also achieved an accuracy of 99%.

Based on the models developed using various scenarios, the set of 19 miRNA features demonstrated the best overall performance. Therefore, the implementation for obtaining these 19 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 will then be matched against the selected 19 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 Lasso + RFE, edgeR, and ROC 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 99%, with a precision of 99%, and recall and F1-score of 100%, 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 89%. This difference indicates a possible variation in miRNA expression patterns across races, which should be considered when developing more inclusive and representative classification models.

## Acknowledgement

## References

- [1] L. Wang, B. Lu, M. He, Y. Wang, Z. Wang, and L. Du, "Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019," *Front Public Health*, vol. 10, Feb. 2022, doi: 10.3389/fpubh.2022.811044.
- [2] A. Yaqoob, R. Musheer Aziz, and N. K. Verma, "Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review," *Human-Centric Intelligent Systems*, vol. 3, no. 4, pp. 588–615, Sep. 2023, doi: 10.1007/s44230-023-00041-3.
- [3] S. Chen *et al.*, "Machine Learning-Based Models Enhance the Prediction of Prostate Cancer," *Front Oncol*, vol. 12, Jul. 2022, doi: 10.3389/fonc.2022.941349.
- [4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput Struct Biotechnol J*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
- [5] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Comput Biol Med*, vol. 121, p. 103761, Jun. 2020, doi: 10.1016/j.compbimed.2020.103761.

- [6] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, “eXtreme Gradient Boosting Algorithm with Machine Learning: a Review,” *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.
- [7] X. Y. Liew, N. Hameed, and J. Clos, “An investigation of XGBoost-based algorithm for breast cancer classification,” *Machine Learning with Applications*, vol. 6, p. 100154, Dec. 2021, doi: 10.1016/j.mlwa.2021.100154.
- [8] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, “An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach,” *MethodsX*, vol. 10, p. 102119, 2023, doi: 10.1016/j.mex.2023.102119.
- [9] X. Guan *et al.*, “Construction of the XGBoost model for early lung cancer prediction based on metabolic indices,” *BMC Med Inform Decis Mak*, vol. 23, no. 1, p. 107, Jun. 2023, doi: 10.1186/s12911-023-02171-x.
- [10] A. Ogunleye and Q.-G. Wang, “XGBoost Model for Chronic Kidney Disease Diagnosis,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.
- [11] T. X. Lu and M. E. Rothenberg, “MicroRNA,” *Journal of Allergy and Clinical Immunology*, vol. 141, no. 4, pp. 1202–1207, Apr. 2018, doi: 10.1016/j.jaci.2017.08.034.
- [12] K. B. Reddy, “MicroRNA (miRNA) in cancer,” *Cancer Cell Int*, vol. 15, no. 1, p. 38, Dec. 2015, doi: 10.1186/s12935-015-0185-1.
- [13] B. Smolarz, A. Durczyński, H. Romanowicz, K. Szyłło, and P. Hogendorf, “miRNAs in Cancer (Review of Literature),” *Int J Mol Sci*, vol. 23, no. 5, p. 2805, Mar. 2022, doi: 10.3390/ijms23052805.
- [14] H. H. Wu, S. Leng, C. Sergi, and R. Leng, “How MicroRNAs Command the Battle against Cancer,” *Int J Mol Sci*, vol. 25, no. 11, p. 5865, May 2024, doi: 10.3390/ijms25115865.
- [15] K. Rishabh, S. Khadilkar, A. Kumar, I. Kalra, A. P. Kumar, and A. B. Kunnumakkara, “MicroRNAs as Modulators of Oral Tumorigenesis—A Focused Review,” *Int J Mol Sci*, vol. 22, no. 5, p. 2561, Mar. 2021, doi: 10.3390/ijms22052561.
- [16] S. Komatsu, H. Kitai, and H. I. Suzuki, “Network Regulation of microRNA Biogenesis and Target Interaction,” *Cells*, vol. 12, no. 2, p. 306, Jan. 2023, doi: 10.3390/cells12020306.
- [17] J. O’Brien, H. Hayder, Y. Zayed, and C. Peng, “Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation,” *Front Endocrinol (Lausanne)*, vol. 9, Aug. 2018, doi: 10.3389/fendo.2018.00402.
- [18] W. A. Haseltine and R. Patarca, “The RNA Revolution in the Central Molecular Biology Dogma Evolution,” *Int J Mol Sci*, vol. 25, no. 23, p. 12695, Nov. 2024, doi: 10.3390/ijms252312695.

- [19] Z. Ali Syeda, S. S. S. Langden, C. Munkhzul, M. Lee, and S. J. Song, "Regulatory Mechanism of MicroRNA Expression in Cancer," *Int J Mol Sci*, vol. 21, no. 5, p. 1723, Mar. 2020, doi: 10.3390/ijms21051723.
- [20] D. Lowder *et al.*, "Racial disparities in prostate cancer: A complex interplay between socioeconomic inequities and genomics," *Cancer Lett*, vol. 531, pp. 71–82, Apr. 2022, doi: 10.1016/j.canlet.2022.01.028.
- [21] S. A. Bigler, C. R. Pound, and X. Zhou, "A Retrospective Study on Pathologic Features and Racial Disparities in Prostate Cancer," *Prostate Cancer*, vol. 2011, pp. 1–7, 2011, doi: 10.1155/2011/239460.
- [22] Z. Ning, S. Yu, Y. Zhao, X. Sun, H. Wu, and X. Yu, "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*, vol. 12, Mar. 2021, doi: 10.3389/fgene.2021.656526.
- [23] F. Bergez-Hernández *et al.*, "Expression Analysis of miRNAs and Their Potential Role as Biomarkers for Prostate Cancer Detection," *Am J Mens Health*, vol. 16, no. 5, Sep. 2022, doi: 10.1177/15579883221120989.
- [24] M. J. Goldman *et al.*, "Visualizing and interpreting cancer genomics data via the Xena platform," *Nat Biotechnol*, vol. 38, no. 6, pp. 675–678, Jun. 2020, doi: 10.1038/s41587-020-0546-8.
- [25] A. Edge, S. and Byrd, D.R. and Compton, C.C. and Fritz, A.G. and Greene, F.L. and Trotti, *AJCC Cancer Staging Manual*. 2010.
- [26] I. Unal, "Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach," *Comput Math Methods Med*, vol. 2017, pp. 1–14, 2017, doi: 10.1155/2017/3762651.
- [27] M. Casarrubios *et al.*, "Tumor microenvironment gene expression profiles associated to complete pathological response and disease progression in resectable NSCLC patients treated with neoadjuvant chemoimmunotherapy," *J Immunother Cancer*, vol. 10, no. 9, p. e005320, Sep. 2022, doi: 10.1136/jitc-2022-005320.
- [28] B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA–cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, Mar. 2013, doi: 10.1093/bioinformatics/btt014.
- [29] Z. Li *et al.*, "LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations," *Nucleic Acids Res*, vol. 51, no. D1, pp. D186–D191, Jan. 2023, doi: 10.1093/nar/gkac999.
- [30] Y. Chen and X. Wang, "miRDB: an online database for prediction of functional microRNA targets," *Nucleic Acids Res*, vol. 48, no. D1, pp. D127–D131, Jan. 2020, doi: 10.1093/nar/gkz757.