Open Access

OBM Genetics

LIDSEN Publishing Inc.

Original Research

A Machine Learning-Based Diagnostic Model for Prostate Cancer Using Circulating MicroRNA Expression Profiles

Minh Trong Quang [1, 2], Minh Nam Nguyen [3, *]

1. Department of Genetics, Faculty of Biology and Biotechnology, University of Science, Vietnam National University Ho Chi Minh City (VNU-HCM), 227 Nguyen Van Cu Street, Cho Quan Ward, Ho Chi Minh City, Vietnam; E-Mail: gtminh@ump.edu.vn
2. School of Pharmacy, University of Medicine and Pharmacy at Ho Chi Minh City, 41-43 Dinh Tien Hoang Street, Sai Gon Ward, Ho Chi Minh City, Vietnam
3. Department of Biomedical Engineering, Faculty of Medicine, University of Health Sciences, Vietnam National University, Ho Chi Minh City, Administrative Building YA1, Hai Thuong Lan Ong Street, VNU-HCM Urban Area, Dong Hoa Ward, Ho Chi Minh City, Vietnam; E-Mail: nmnam@uhsvnu.edu.vn

* **Correspondence:** Minh Nam Nguyen; E-Mail: nmnam@uhsvnu.edu.vn

Academic Editor: Xuehuo Zeng

OBM Genetics
2026, volume 10, issue 3
doi:10.21926/obm.genet.2603346

Received: March 10, 2026
Accepted: June 22, 2026
Published: July 01, 2026

Abstract

Prostate cancer (PCa) is one of the most common malignancies among men worldwide, and early detection is critical for improving clinical outcomes. Circulating microRNAs (miRNAs) have emerged as promising non-invasive biomarkers for cancer diagnosis due to their stability in blood and association with tumor-related molecular alterations. In this study, machine learning (ML) methods were applied to large-scale circulating miRNA expression data to develop a diagnostic model for PCa detection. Serum miRNA expression profiles were obtained from the Gene Expression Omnibus dataset GSE211692, which included 6,920 samples comprising 1,027 PCa cases and 5,893 non-cancer controls. To reduce the risk of overfitting and information leakage, preprocessing, normalization, feature selection, and hyperparameter optimization were performed within the training and cross-validation framework, with the held-out testing set used only for final internal evaluation. Four ML

algorithms, namely Logistic Regression, K-Nearest Neighbors, Random Forest, and CatBoost, were implemented. Principal component analysis (PCA) was additionally performed on both the training and held-out test datasets to visualize the sample distribution by case-control status. Although PCA showed clear separation between PCa and non-cancer samples, complete batch-related metadata were unavailable; therefore, potential technical batch effects could not be fully excluded. Among the evaluated algorithms, the Random Forest classifier showed the strongest internal diagnostic performance. A three-miRNA panel comprising miR-1290, miR-1307-3p, and miR-4783-3p demonstrated strong discriminatory capability for PCa classification. However, because the model was developed and evaluated on a single public dataset without external validation, the reported performance should be interpreted with caution. Further validation in independent clinical cohorts, comparison with established biomarkers such as PSA, and integration of clinicopathological variables are required before clinical translation can be considered.

Keywords
Prostate cancer; circulating microRNA; machine learning; random forest; liquid biopsy; biomarker; diagnostic model; batch effect; information leakage; internal validation

1. Introduction

Prostate cancer (PCa) is one of the most frequently diagnosed malignancies among men worldwide and remains a major public health challenge. According to recent global cancer statistics, PCa accounted for approximately 7.3% of all newly diagnosed cancers worldwide in 2022, making it one of the most common cancers globally and one of the leading malignancies in men [1, 2]. The burden of PCa is expected to increase substantially in the coming decades due to population aging, longer life expectancy, and improved cancer detection in many regions [3]. Therefore, improving early detection strategies remains a key priority for reducing disease-related morbidity and mortality.

Early diagnosis is particularly important because clinical outcomes differ markedly between localized and advanced PCa. Patients with localized disease generally have favorable survival outcomes, whereas metastatic or treatment-resistant disease is associated with poorer prognosis and more limited therapeutic options. In clinical practice, prostate-specific antigen (PSA) testing remains the most widely used blood-based tool for PCa screening and monitoring. However, PSA has important limitations, particularly its limited tumor specificity. Elevated PSA levels may occur in benign prostatic hyperplasia, prostatitis, and other non-malignant conditions, which can lead to false-positive findings, unnecessary biopsies, overdiagnosis, and overtreatment [4, 5]. Recent reviews of blood- and urine-based biomarkers in PCa have emphasized that, although PSA remains central in clinical practice, its limited specificity and sensitivity underscore the need for complementary biomarkers to improve diagnostic accuracy and clinical decision-making [6].

Liquid biopsy has emerged as a promising approach for non-invasive cancer detection and monitoring. Compared with tissue biopsy, liquid biopsy can provide molecular information using minimally invasive specimens such as blood, serum, plasma, or urine. Among liquid biopsy-derived

biomarkers, circulating microRNAs (miRNAs) have attracted substantial attention. MiRNAs are small non-coding RNA molecules, approximately 18-25 nucleotides in length, that regulate gene expression at the post-transcriptional level by inhibiting translation or promoting degradation of target messenger RNAs [7, 8]. Through these regulatory functions, miRNAs participate in key biological processes including cell proliferation, apoptosis, differentiation, angiogenesis, invasion, and metastasis. Dysregulation of miRNA expression has been widely associated with cancer initiation and progression [9, 10]. Circulating miRNAs are particularly attractive biomarker candidates because they are stable in biological fluids and can be protected from degradation through association with extracellular vesicles, protein complexes, or lipoproteins [11]. These properties make circulating miRNAs suitable for liquid biopsy-based diagnostic applications. In PCa, several studies have reported altered circulating miRNA profiles in patients compared with non-cancer controls, suggesting their potential value as non-invasive diagnostic, prognostic, and monitoring biomarkers [12, 13]. Recent literature on genitourinary cancers has further highlighted circulating miRNAs as promising early-detection biomarkers, while also emphasizing the need for analytical standardization, independent validation, and careful evaluation of clinical utility before routine implementation [14].

Despite their promise, the diagnostic performance of individual circulating miRNAs can vary across studies due to biological heterogeneity, differences in sample types, RNA extraction methods, measurement platforms, normalization strategies, and cohort composition. Multi-marker signatures provide more robust diagnostic performance than single biomarkers. Machine learning (ML) methods provide a useful analytical framework for identifying biomarker signatures from high-dimensional molecular datasets. ML algorithms can capture complex relationships among multiple molecular features and may improve disease classification when applied appropriately [15, 16]. In cancer research, ML has been increasingly used to analyze genomic, transcriptomic, imaging, and liquid biopsy data for diagnosis, prognosis, and risk stratification [17]. However, omics-based ML studies require careful methodological design. Extremely high classification performance in high-dimensional molecular datasets should be interpreted with caution, as it may reflect overfitting, information leakage, or technical confounding rather than true biological discrimination. Information leakage can occur when preprocessing, normalization, feature selection, or hyperparameter tuning is performed before data splitting or outside the cross-validation framework, allowing information from validation or testing samples to influence model development [18, 19]. In addition, batch effects are a major concern in high-throughput molecular datasets. Technical variation related to sample processing, experimental batch, platform, or data acquisition can create artificial separation between groups, particularly when batch structure is correlated with case-control status [20]. Therefore, robust ML-based biomarker studies should include leakage-controlled analytical workflows and assessment of potential batch effects.

Based on these considerations, the present study aimed to develop an ML-based diagnostic model that uses serum circulating miRNA expression profiles to classify PCa. Publicly available data from the Gene Expression Omnibus (GEO) were analyzed to identify a compact circulating miRNA panel with potential diagnostic relevance. To address concerns regarding information leakage and potential batch-effect confounding, the analytical framework was revised to ensure that preprocessing, normalization, feature selection, and hyperparameter optimization were performed during training and cross-validation. The held-out testing set was reserved solely for final internal evaluation. Principal component analysis (PCA) was additionally performed on both the training and

held-out test datasets to visualize the global distribution of samples by PCa status. Although PCA showed clear separation between PCa and non-cancer samples in both subsets, complete batch-related metadata were not available; therefore, clustering by experimental batch or sample-processing identifiers could not be assessed directly. Because no independent external validation cohort was available in the present analysis, the results are interpreted as internal evidence only, and external validation remains necessary before clinical translation can be considered.

2. Materials and Methods

2.1 Study Design and Data Acquisition

Publicly available datasets describing circulating miRNA expression profiles associated with PCa were retrieved from the GEO database maintained by the National Center for Biotechnology Information (NCBI). GEO is a public repository that archives high-throughput gene expression data, including microarray- and sequencing-based datasets, and enables secondary analysis of publicly available transcriptomic data [21]. The database was searched using the following query: [“prostate neoplasms” (MeSH Terms) OR “prostate cancer” (All Fields)] AND “Homo sapiens” (porgn) AND [“microRNAs” (MeSH Terms) OR “miRNA” (All Fields)].

To ensure biological relevance and methodological consistency, retrieved datasets were screened according to predefined eligibility criteria. Datasets were included if they investigated circulating miRNA expression profiles in patients with PCa, clearly described the experimental platform used to quantify miRNA expression, and included serum, plasma, or whole-blood control samples obtained from individuals without PCa or other malignancies. Datasets were excluded if they were generated from cancer cell lines, animal models, or human tumor xenografts, or if blood samples were collected from patients who had undergone surgery, chemotherapy, radiotherapy, or hormone therapy.

After screening, the GSE211692 dataset was selected for model development. This dataset contains serum circulating miRNA expression profiles generated using the 3D-Gene Human miRNA V21 platform,comprising 6,920 samples: 1,027 PCa cases and 5,893 non-cancer controls. Because no suitable independent GEO dataset with comparable circulating miRNA profiles, compatible sample type, and complete availability of the final selected miRNAs was available for external validation, the present analysis was restricted to internal model development and internal testing. Therefore, the results should be interpreted as internally validated findings rather than evidence of external generalizability.

2.2 Data Preprocessing

The collected miRNA expression dataset was preprocessed before ML analysis to ensure data quality, comparability, and reproducibility. Expression values from the original GEO dataset were used for downstream analysis. Missing values, if present, were handled using median imputation. To reduce the risk of information leakage, imputation parameters were estimated from the training data alone and were not derived from the held-out test dataset.

The dataset was split into a training set and an internal held-out test set at an 80:20 ratio. Stratified sampling was applied to preserve the proportion of PCa and non-PCa samples in both subsets. The training dataset was used for preprocessing parameter estimation, feature selection,

model development, cross-validation, and hyperparameter optimization. The held-out testing dataset was kept separate and used only once for the final internal performance evaluation.

Feature scaling was performed using z-score standardization with the StandardScaler function. Scaling parameters were fitted using the training data and then applied to the held-out testing data. During cross-validation, imputation and scaling were performed within each training fold, and the fitted parameters were applied only to the corresponding validation fold. This fold-specific preprocessing strategy was used to prevent information from validation or testing samples from influencing model development.

Because the dataset was imbalanced, with more non-cancer controls than PCa cases, class imbalance was considered during model construction. For algorithms that supported class weighting, balanced class-weighting strategies were evaluated during model optimization. In addition to accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and AUC, these metrics were reported to provide a more complete assessment of model performance in the presence of class imbalance.

2.3 Feature Selection

Feature selection was performed to identify circulating miRNAs with strong discriminatory ability for PCa classification. Differential expression analysis was conducted using the limma package in R, which applies linear modeling and empirical Bayes methods for gene expression analysis [22]. MiRNAs were considered significantly differentially expressed when they met both criteria: $|\log_2$ fold change$| \geq 3$ and false discovery rate-adjusted p-value ≤ 0.001. The false discovery rate was controlled using the Benjamini-Hochberg correction method [23]. To mitigate the risk of information leakage, feature selection was incorporated into the training framework rather than applied to the entire dataset prior to model evaluation. During cross-validation, differential expression filtering and recursive feature addition were performed using only the training portion of each fold. The corresponding validation fold was not used during biomarker selection.

After the initial differential expression filtering, recursive feature addition was applied to refine the feature set and identify a compact miRNA panel. Features were sequentially added to the model, and their contributions to classification performance were evaluated within a cross-validation framework. The final miRNA panel was selected based on consistent discrimination across training folds and was locked before evaluation on the held-out test dataset. No additional feature reselection was performed on the held-out test data.

2.4 Machine-Learning Model Development

Four supervised ML algorithms were implemented to construct diagnostic models based on circulating miRNA expression profiles: Logistic Regression, K-Nearest Neighbors, Random Forest, and CatBoost. Logistic Regression was used as a baseline linear classifier due to its interpretability and widespread use in biomedical prediction modeling. K-Nearest Neighbors classifies samples based on similarity to neighboring samples in the feature space [24]. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection, thereby improving robustness and reducing overfitting compared with individual decision trees [25]. CatBoost is a gradient boosting algorithm designed to improve predictive performance and reduce prediction bias in structured datasets [26].

Model development was performed within a cross-validation framework, using only the training dataset. For each cross-validation fold, preprocessing, feature selection, model fitting, and validation were conducted independently. The validation fold was not used for imputation, scaling, feature selection, or model training. This workflow was designed to reduce information leakage and provide a more realistic estimate of internal model performance.

Hyperparameter optimization was performed using GridSearchCV combined with five-fold cross-validation. For each algorithm, predefined hyperparameter combinations were evaluated, and the configuration with the best cross-validation performance was selected. The held-out testing dataset was not used during hyperparameter tuning. After the optimal hyperparameters and final miRNA panel were selected, the final model was retrained using the full training dataset and then applied to the held-out testing dataset for final internal evaluation.

2.5 Principal Component Analysis and Exploratory Assessment of Data Structure

Because high-throughput molecular datasets may be affected by technical variation, PCA was performed as an exploratory quality-control step before the final interpretation of model performance. PCA was conducted separately for the training and held-out testing datasets using the circulating miRNA expression matrix. Samples were colored by PCa status to assess whether global expression patterns separated PCa from non-cancer samples.

This analysis was used to evaluate whether the case-control separation observed in the ML analysis was also reflected in the overall expression structure of the training and testing datasets. Because complete batch-related metadata were not available in GSE211692, PCA could not be used to directly assess clustering according to experimental batch, sample-processing group, or other technical identifiers. Therefore, PCA was interpreted as an exploratory assessment of global data structure rather than a definitive batch-effect analysis. The possibility of batch-related confounding could not be fully excluded and was taken into account when interpreting the internal model's performance.

2.6 Evaluation of Model Performance

Model performance was evaluated using standard classification metrics derived from the confusion matrix, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and F1-score. Sensitivity was defined as the proportion of true PCa cases correctly identified by the model, and specificity as the proportion of non-cancer controls correctly classified as negative. PPV represented the proportion of predicted positive samples that were true PCa cases, while NPV represented the proportion of predicted negative samples that were true controls. Accuracy measured the overall proportion of correctly classified samples, and the F1-score was calculated as the harmonic mean of precision and sensitivity.

The area under the receiver operating characteristic curve (AUC) was used as the primary measure of discriminatory ability [27]. Model performance was first estimated using five-fold cross-validation within the training dataset and then evaluated on the held-out internal test dataset. Internal testing performance was reported separately from cross-validation performance. Because no independent external validation dataset was available, the reported results reflect only internal validation. The overall analytical workflow of the study is summarized in Figure 1.

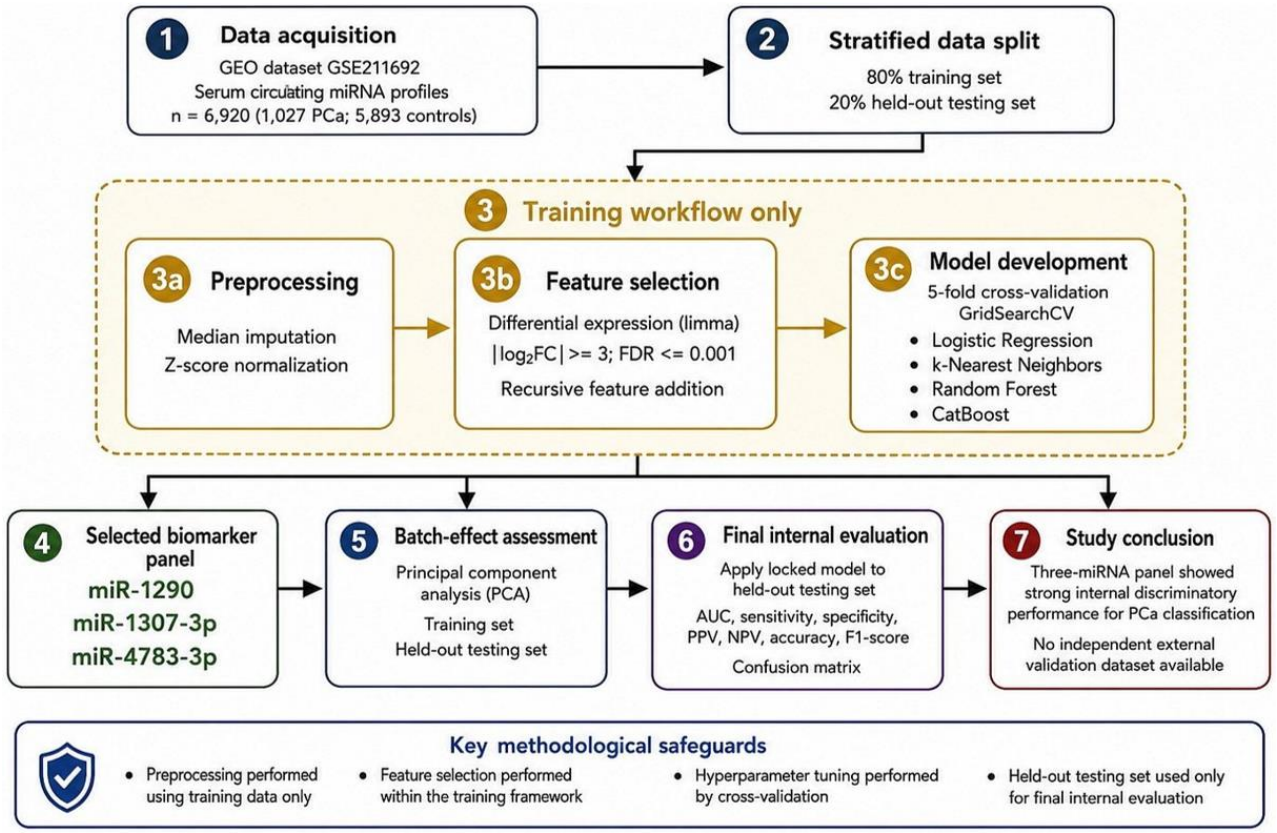

Figure 1 Analytical workflow of the study.

2.7 Software and Computational Environment

All analyses were performed using Python and R. Differential expression analysis was conducted in R using the limma package [22]. ML model development, preprocessing, cross-validation, hyperparameter optimization, and performance evaluation were implemented in Python using scikit-learn [28]. CatBoost models were implemented using the official CatBoost Python package [26]. Data manipulation and visualization were performed using NumPy, pandas, Matplotlib, and Seaborn. PCA was performed and visualized using Python-based analytical packages.

3. Results

3.1 Characteristics of the Dataset

After systematic retrieval and screening of circulating miRNA expression datasets from the GEO database, GSE211692 was selected for further analysis. This dataset was generated using the 3D-Gene Human miRNA V21_1.0.0 microarray platform and contains large-scale serum circulating miRNA expression profiles. The dataset comprised 6,920 serum samples, including 1,027 PCa samples and 5,893 non-cancer control samples. Table 1 summarizes the dataset's overall characteristics used in this study.

Table 1 Characteristics of the study dataset.

Dataset	Total samples	Prostate cancer	Non-cancer controls	Platform
GSE211692	6920	1027	5893	3D-Gene human miRNA V21

Because the dataset contained substantially more non-cancer controls than PCa cases, the class distribution was imbalanced. Therefore, stratified data splitting was applied to preserve the case-control proportion in both the training and internal testing datasets. In addition, model performance was evaluated using multiple metrics, including sensitivity, specificity, PPV, NPV, F1-score, accuracy, and AUC, rather than accuracy alone.

The dataset was randomly split into a training set and an internal held-out test set at an 80:20 ratio. The training dataset consisted of 822 PCa samples and 4,714 control samples, whereas the held-out testing dataset included 205 PCa samples and 1,179 control samples. The held-out testing set was not used during preprocessing parameter estimation, feature selection, hyperparameter tuning, or model training.

3.2 Principal Component Analysis of the Training and Testing Datasets

PCA was performed to visualize the global distribution of serum circulating miRNA expression profiles in the training and held-out testing datasets. As shown in Figure 2A, PCA of the training dataset clearly separated PCa and non-cancer samples, primarily along the first principal component. In the training dataset, PC1 accounted for 75.22% of the total variance, whereas PC2 accounted for 12.93%. This pattern suggests that the major source of variation in the training data was strongly associated with case-control status. A similar PCA pattern was observed in the held-out testing dataset. As shown in Figure 2B, PCa and non-cancer samples also showed clear separation, although the variance distribution differed from that of the training set. In the test dataset, PC1 accounted for 45.04% of the total variance, and PC2 accounted for 32.45%. Together, these two components accounted for 77.49% of the total variance, indicating that the testing samples retained a strong case-control expression structure.

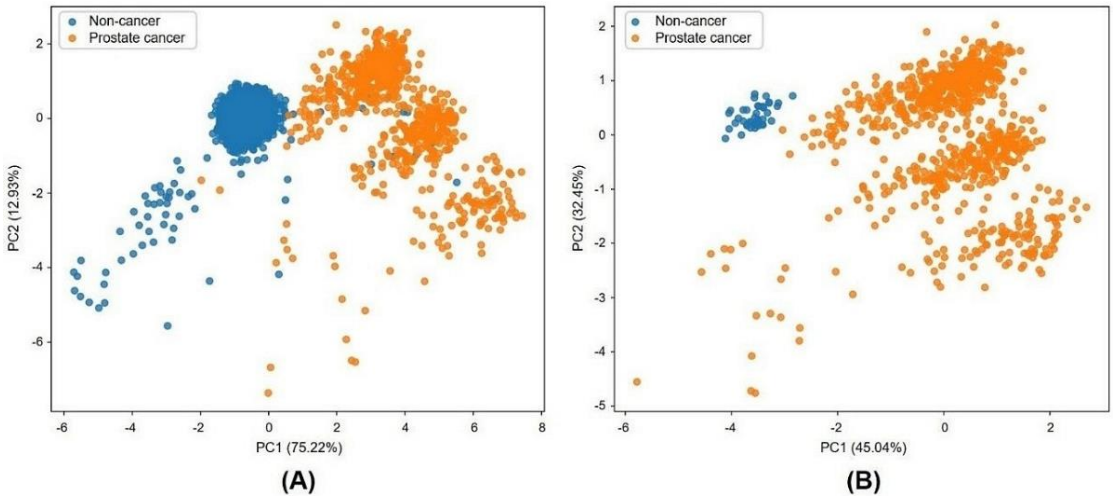

Figure 2 Principal component analysis of serum circulating miRNA expression profiles. (A) Training dataset. (B) Held-out testing dataset. Samples are colored by disease status: non-cancer controls and prostate cancer.

Overall, PCA showed consistent separation between PCa and non-cancer samples in both the training and held-out testing datasets. These findings support the presence of a strong discriminative expression pattern in the analyzed dataset. However, because complete batch-related metadata were not available, PCA could not definitively determine whether technical batch effects contributed to the observed separation. Therefore, although the PCA results support strong internal case-control discrimination, the possibility of batch-related confounding cannot be ruled out entirely.

3.3 Identification of the Optimal Circulating miRNA Biomarker Panel

Feature selection identified three circulating miRNAs with strong discriminatory ability for PCa classification: miR-1290, miR-1307-3p, and miR-4783-3p. These miRNAs were selected as the optimal biomarker panel for subsequent ML model development.

All three miRNAs showed higher expression levels in serum samples from patients with PCa than in non-cancer controls. The expression distributions of miR-1290, miR-1307-3p, and miR-4783-3p demonstrated clear differences between the two groups, supporting their potential contribution to PCa classification (Figure 3). However, because feature selection in high-dimensional molecular datasets can strongly influence model performance, the final three-miRNA panel should be interpreted as a candidate biomarker signature requiring further validation. Independent external cohorts will be necessary to determine whether this panel remains stable across different populations, platforms, and sample-processing conditions.

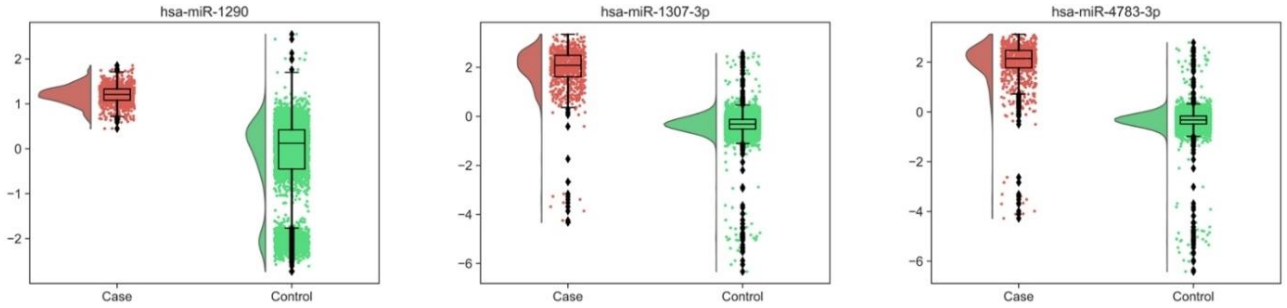

Figure 3 Distribution of circulating miR-1290, miR-1307-3p, and miR-4783-3p expression levels in PCa and non-cancer control samples.

3.4 Performance of Machine-Learning Models

A total of 28 ML models were developed using different algorithms and feature combinations. Among these models, seven demonstrated high internal diagnostic performance. The evaluated models were constructed using different combinations of the three selected circulating miRNAs, including miR-1290, miR-1307-3p, and miR-4783-3p.

The best-performing models are summarized in Table 2. Both CatBoost and Random Forest showed strong internal classification performance. Among the evaluated models, the Random Forest classifier using the combined three-miRNA panel achieved the highest internal test performance. Importantly, the performance values reported here represent internal testing results derived from a held-out subset of the same source dataset. Therefore, these values should not be interpreted as evidence of external clinical generalizability.

Table 2 Best-performing machine-learning models for prostate cancer diagnosis.

Model	Algorithm	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)	Accuracy (%)	AUC (%)
1	CatBoost	95.61	97.54	87.11	99.22	91.16	97.25	99.50
2	CatBoost	97.56	98.05	89.69	99.57	93.46	97.98	99.22
3	Random Forest	97.07	97.88	88.84	99.48	92.77	97.76	98.81
4	CatBoost	98.54	99.41	96.65	99.74	97.58	99.28	99.96
5	CatBoost	99.02	99.49	97.13	99.83	98.07	99.42	99.98
6	Random Forest	98.05	98.64	92.63	99.66	95.26	98.55	99.55
7	Random Forest	99.51	99.58	97.61	99.91	98.55	99.57	99.98

Note: Model 1: hsa-miR-1290; Model 2: hsa-miR-1307-3p; Model 3: hsa-miR-4783-3p; Model 4: hsa-miR-1290 + hsa-miR-1307-3p; Model 5: hsa-miR-1290 + hsa-miR-4783-3p; Model 6: hsa-miR-1307-3p + hsa-miR-4783-3p; Model 7: hsa-miR-1290 + hsa-miR-1307-3p + hsa-miR-4783-3p. All performance metrics were calculated on the independent testing dataset.

The Random Forest model using the three-miRNA panel achieved a sensitivity of 99.51%, specificity of 99.58%, PPV of 97.61%, NPV of 99.91%, F1-score of 98.55%, accuracy of 99.57%, and AUC of 99.98% on the internal held-out testing dataset. While these results indicate strong internal discriminatory performance, the extremely high AUC should be interpreted cautiously because omics-based machine-learning models may be affected by overfitting, information leakage, or technical confounding if the analytical workflow is not carefully controlled.

3.5 Roc Analysis and Confusion Matrix of the Optimal Model

ROC analysis was performed to further evaluate the optimal model's discriminatory ability (Figure 4). The Random Forest classifier using miR-1290, miR-1307-3p, and miR-4783-3p achieved an AUC of 0.9998 in the held-out internal test set. The ROC curve remained close to the upper-left corner of the plot, indicating high sensitivity and specificity across classification thresholds.

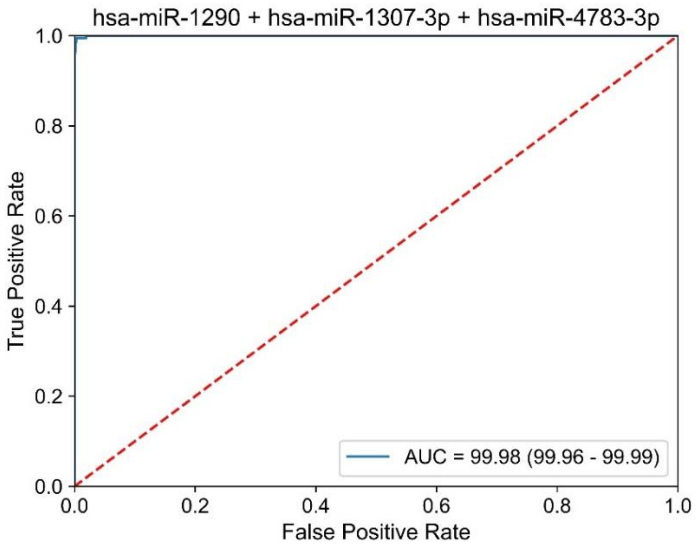

Figure 4 ROC curve of the optimal Random Forest model using the three-miRNA panel for PCa classification.

The classification performance of the optimal model was further examined using a confusion matrix (Figure 5). The majority of samples were correctly classified, with only a small number of false-positive and false-negative predictions. These results support the strong internal performance of the three-miRNA Random Forest model. Nevertheless, because the testing samples were derived from the same GEO dataset as the training samples, the ROC and confusion matrix results should be considered internal validation only. External validation in an independent dataset remains necessary to assess whether the model maintains performance across different cohorts and experimental settings.

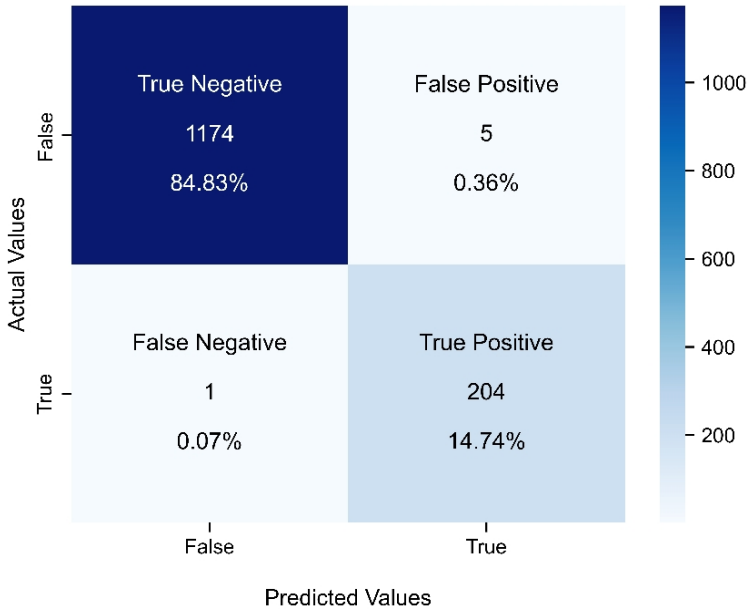

Figure 5 Confusion matrix of the optimal Random Forest diagnostic model in the held-out internal testing dataset.

4. Discussion

In this study, circulating miRNA expression profiles were integrated with ML algorithms to develop a diagnostic model for PCa classification. Four supervised learning algorithms, including Logistic Regression, K-Nearest Neighbors, Random Forest, and CatBoost, were evaluated using serum circulating miRNA expression data from GSE211692. A compact three-miRNA panel comprising miR-1290, miR-1307-3p, and miR-4783-3p showed strong discriminatory potential between PCa cases and non-cancer controls. Among the evaluated algorithms, the Random Forest classifier using this three-miRNA panel achieved the strongest performance in internal testing.

The principal finding of this study is that a small circulating miRNA signature may contain substantial diagnostic information for distinguishing PCa samples from non-cancer controls. This is consistent with the broader concept that circulating miRNAs can reflect tumor-associated molecular alterations and may serve as non-invasive biomarkers in liquid biopsy-based cancer detection. Recent reviews have also emphasized the potential of circulating miRNAs in genitourinary cancers, while highlighting the need for careful analytical validation, technical standardization, and independent clinical confirmation before clinical use [14]. However, the extremely high internal performance observed in this study should be interpreted cautiously. In omics-based ML studies, very high AUC values may reflect a true biological signal, but they may also arise from overfitting,

information leakage, or batch-effect confounding. Information leakage can occur when preprocessing, normalization, feature selection, or hyperparameter tuning is performed using information from validation or testing samples. Such leakage can lead to overly optimistic performance estimates and may limit reproducibility across external cohorts [18, 19]. To reduce this risk, the analytical framework ensured that imputation, scaling, feature selection, model training, and hyperparameter optimization were performed within the training and cross-validation procedures. At the same time, the held-out test set was reserved solely for final internal evaluation.

Batch effects represent another important concern in high-throughput molecular datasets. Technical variation related to sample collection, RNA extraction, experimental batch, array processing, or data normalization can create artificial differences between groups. This is particularly problematic when the technical structure correlates with case-control status, because an ML model may learn batch-related differences rather than disease-associated biological signals [20]. To explore the dataset's global structure, PCA was performed separately on the training and held-out test datasets. The PCA plots showed clear separation between PCa and non-cancer samples in both subsets, suggesting that the case-control signal was preserved after data splitting. Nevertheless, because complete batch-related metadata were unavailable, PCA could not directly assess clustering according to experimental batch or sample-processing identifiers. Therefore, the possibility of batch-effect confounding cannot be fully excluded, and the model performance should be interpreted as internal evidence that requires confirmation in independent external cohorts.

The three miRNAs identified in this study have potential biological relevance in PCa and cancer progression. MiR-1290 has been reported as an oncogenic miRNA in several malignancies and has been associated with tumor progression, aggressive disease behavior, and poor clinical outcomes. In PCa, circulating or exosomal miR-1290 has been linked to advanced disease and poor prognosis, suggesting that its presence in the blood may reflect tumor-associated molecular activity [29]. MiR-1307-3p has also been implicated in cancer-related proliferation pathways. Previous experimental evidence suggests that miR-1307 can promote PCa cell proliferation by targeting FOXO3A, a tumor-suppressive transcription factor involved in apoptosis, oxidative stress response, and cell-cycle regulation [30]. Compared with miR-1290 and miR-1307-3p, miR-4783-3p remains less extensively characterized in PCa. Nevertheless, integrated bioinformatics analyses suggest that miR-4783-3p may regulate target genes and pathways relevant to PCa biology [9].

Although these biological observations support the plausibility of the selected miRNA panel, the functional interpretation remains hypothesis-generating. The present study did not experimentally validate the mechanistic roles of miR-1290, miR-1307-3p, or miR-4783-3p in PCa cell models or patient-derived specimens. Therefore, future studies should investigate whether these miRNAs directly regulate PCa-associated pathways, including cell proliferation, apoptosis, angiogenesis, epithelial-mesenchymal transition, androgen receptor signaling, metastatic progression, and treatment resistance. Functional validation would strengthen the biological and translational relevance of the proposed biomarker panel [31-33].

The proposed circulating miRNA model may have potential value as a complementary diagnostic tool, but it should not be considered a replacement for established clinical biomarkers at this stage. PSA remains the most widely used blood-based biomarker for PCa screening and monitoring. However, PSA has limited tumor specificity, and elevated PSA levels may occur in benign prostatic hyperplasia, prostatitis, and other non-malignant conditions. These limitations can contribute to unnecessary biopsies, overdiagnosis, and overtreatment. Recent reviews of blood- and urine-based

PCa biomarkers have emphasized that emerging markers, including PHI, 4Kscore, PCA3, SelectMDx, extracellular vesicle-associated biomarkers, and circulating miRNAs, may improve diagnostic precision or complement PSA-based pathways. Still, they require rigorous validation before broad clinical implementation [6].

A direct comparison between the proposed three-miRNA model and PSA, PSA derivatives, prostate MRI, or other established diagnostic approaches was not possible in the present study because matched clinical variables were not available in the GEO dataset analyzed. Therefore, the model's clinical utility remains uncertain. Future validation cohorts should include PSA concentration, prostate volume, digital rectal examination findings, MRI results, Gleason score, tumor stage, and disease risk group. Such data would enable the determination of whether the miRNA panel provides incremental diagnostic value beyond currently used clinical tools.

This study has several strengths. First, the analysis used a large public serum miRNA dataset, which provided sufficient sample size for internal model development and testing. Second, multiple supervised ML algorithms were evaluated, allowing comparison across different model types. Third, the final biomarker panel was compact, consisting of only three circulating miRNAs, which may be more feasible for future assay development than larger molecular signatures. Fourth, the framework explicitly addresses methodological concerns raised for omics-based ML studies by emphasizing leakage-controlled preprocessing, fold-specific feature selection, and PCA-based assessment of batch effects. These additions improve transparency and help clarify the conditions under which the internal performance estimates should be interpreted.

Despite these strengths, several limitations must be considered. First, no independent external GEO dataset suitable for direct validation of the final three-miRNA panel was available for this analysis. Therefore, the model was evaluated solely through internal testing using a held-out subset of the same source dataset. This limits conclusions about generalizability across independent cohorts, different populations, alternative sample-processing protocols, and other miRNA profiling platforms. Second, although internal performance was high, the possibility of overfitting or optimistic performance estimation cannot be fully ruled out. Third, batch-effect assessment was limited by the availability and completeness of technical annotation in the public dataset. PCA can reveal global clustering patterns, but it cannot fully replace experimental batch control or prospective validation. Fourth, important clinicopathological variables, including age, PSA level, Gleason score, tumor stage, disease risk category, prostate volume, and treatment history, were not available for all samples. Fifth, the biological functions of the selected miRNAs were inferred from previous literature and bioinformatics evidence rather than experimentally validated in this study. Given these limitations, the findings should be interpreted as preliminary internal evidence supporting the potential diagnostic relevance of the three-miRNA panel rather than definitive evidence of clinical diagnostic accuracy. The very high internal AUC should not be described as near-perfect clinical performance. Instead, it should be considered a signal that requires confirmation using independent external datasets and prospective clinical cohorts. Future studies should apply a fully leakage-controlled pipeline, report batch-effect analyses, validate the locked model without re-selecting features, and compare the miRNA panel directly with PSA-based and imaging-based diagnostic pathways.

5. Conclusions

In conclusion, this study suggests that serum miRNA expression profiles, combined with ML methods, may provide a promising framework for non-invasive PCa classification. A compact panel consisting of miR-1290, miR-1307-3p, and miR-4783-3p showed strong internal discriminatory potential, and the Random Forest classifier achieved the best performance among the evaluated algorithms. However, because the analysis was based on a single public dataset and no external validation cohort was available, the model should be regarded as internally validated and hypothesis-generating. Independent external validation, direct comparison with established clinical biomarkers such as PSA, integration of clinicopathological variables, and functional validation of the selected miRNAs are required before this approach can be considered for clinical translation.

Acknowledgments

Minh Trong Quang was funded by the Master, PhD Scholarship Program of Vingroup Innovation Foundation (VINIF), code VINIF.2021.ThS.69 and VINIF.2022.ThS.054.

Author Contributions

Minh Trong Quang: Conceptualization, methodology, data curation, formal analysis, software, visualization, and preparation of the original draft. Minh Nam Nguyen: Conceptualization, supervision, validation, project administration, writing review, and editing.

Funding

This research was funded by the Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2023-44-06. The funders had no role in the study design, data collection, analysis, or interpretation, writing of the manuscript, or decision to publish the results.

Competing Interests

The authors have declared that no competing interests exist.

Data Availability Statement

The datasets analyzed in this study are publicly available in the Gene Expression Omnibus (GEO) repository of the National Center for Biotechnology Information (NCBI). The dataset used in this study can be accessed under the accession number GSE211692 (https://www.ncbi.nlm.nih.gov/geo/). Additional data supporting the findings of this study are available from the corresponding author upon reasonable request.

AI-Assisted Technologies Statement

The authors declare that AI-assisted technologies were used only to support language editing and improve the clarity of the manuscript. The authors have reviewed and edited the content as necessary and take full responsibility for the accuracy, integrity, and originality of the work.

References

1. Nguyen TA, Duong KC, Do XD, Nguyen AD, Quang MT. Prostate cancer in Asia: Epidemiology, association with human development index and projections to 2040. Asian Pac J Cancer Care. 2026; 11: 353-361.
2. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024; 74: 229-263.
3. James ND, Tannock I, N'Dow J, Feng F, Gillessen S, Ali SA, et al. The lancet commission on prostate cancer: Planning for the surge in cases. Lancet. 2024; 403: 1683-1722.
4. Loeb S, Bjurlin MA, Nicholson J, Tammela TL, Penson DF, Carter HB, et al. Overdiagnosis and overtreatment of prostate cancer. Eur Urol. 2014; 65: 1046-1055.
5. Cornford P, van den Bergh RC, Briers E, Van den Broeck T, Brunckhorst O, Darraugh J, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer-2024 update. Part I: Screening, diagnosis, and local treatment with curative intent. Eur Urol. 2024; 86: 148-163.
6. Crocetto F, Musone M, Chianese S, Conforti P, Selvaggio GD, Caputo VF, et al. Blood and urine-based biomarkers in prostate cancer: Current advances, clinical applications, and future directions. J Liq Biopsy. 2025; 9: 100305.
7. Bartel DP. MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell. 2004; 116: 281-297.
8. Quang MT, Nguyen MN. The potential of microRNAs in cancer diagnostic and therapeutic strategies: A narrative review. J Basic Appl Zool. 2024; 85: 7.
9. Nguyen MT, Quang MT. Integrated bioinformatics analysis of hsa-miR-4783-3p target genes and functions in prostate cancer. Pharm Sci Asia. 2024; 51: 233-240.
10. Quang MT, Nguyen MN, Than VT. The role and regulation of cell death in cancer. Prog Mol Biol Transl Sci. 2025; 217: 135-161.
11. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci. 2008; 105: 10513-10518.
12. Haldrup C, Kosaka N, Ochiya T, Borre M, Høyer S, Orntoft TF, et al. Profiling of circulating microRNAs for prostate cancer biomarker discovery. Drug Deliv Transl Res. 2014; 4: 19-30.
13. Urabe F, Matsuzaki J, Yamamoto Y, Kimura T, Hara T, Ichikawa M, et al. Large-scale circulating microRNA profiling for the liquid biopsy of prostate cancer. Clin Cancer Res. 2019; 25: 3016-3025.
14. Cicatiello AG, Musone M, Imperatore S, Giulioni C, La Rocca R, Cafarelli A, et al. Circulating miRNAs in genitourinary cancer: Pioneering advances in early detection and diagnosis. J Liq Biopsy. 2025; 8: 100296.
15. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015; 16: 321-332.
16. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015; 13: 8-17.
17. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. RadioGraphics. 2017; 37: 505-515.

18. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci. 2002; 99: 6562-6566.
19. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinf. 2006; 7: 91.
20. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11: 733-739.
21. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30: 207-210.
22. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 57: 289-300.
24. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967; 13: 21-27.
25. Breiman L. Random forests. Mach Learn. 2001; 45: 5-32.
26. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018; 31: 6638-6648.
27. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143: 29-36.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011; 12: 2825-2830.
29. Huang X, Yuan T, Liang M, Du M, Xia S, Dittmar R, et al. Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. Eur Urol. 2015; 67: 33-41.
30. Qiu X, Dou Y. miR-1307 promotes the proliferation of prostate cancer by targeting FOXO3A. Biomed Pharmacother. 2017; 88: 430-435.
31. Huynh TT, Nguyen TN, Nguyen TA, Nguyen AD, Quang MT. Convergence of gene therapy and vaccine platforms in the post-pandemic era: A mini review. Trends Sci. 2026; 23: 12197.
32. Nguyen AD, Quang MT. CRISPR/Cas9 genome editing in oncology: Mechanisms, therapeutic platforms and translational challenges. Mol Biotechnol. 2025; 68: 2201-2229.
33. Ha TA, Nguyen TN, Nguyen TA, Huynh TT, Nguyen AD, Quang MT. CRISPR/Cas9 editing of the β-globin locus in sickle cell disease and β-thalassemia: Mechanistic rationale, clinical evidence, and future directions. Trends Sci. 2026; 23: 12982.