ISSN 2085-4552

Prostate Cancer Screening for Specific Races Using Bioinformatics and Artificial Intelligence on Genomic Data

David Agustriawan [1], Marlinda Vasty Overbeek [2], Angga Aditya Permana [3]
[1,2,3] Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia
[1] david.agustriawan@umn.ac.id, [2] marlinda.vasty@umn.ac.id, [3] angga.permana@umn.ac.id

Accepted 07 August 2024
Approved 14 August 2024

***Abstract*— Prostate cancer is one of a deathly cancer worldwide. The higher incidence and mortality rate shows that it is an urgent call for all of us to fight against it in our own way. This study develops an artificial intelligence system to screening prostate cancer from normal patients in a specific race. Gene expression and its phenotype dataset was downloaded from xenabrowser.net. Data preprocessing and filtering based on a particular race, bioinformatics computational analysis to determine the features and machine learning algorithm such as decision tree and random forest are used to develop AI model. All the procedure and analysis was performed using python programming. The result show that only White and Black African American has a proper number of dataset while Asian and American Indian has a very lack dataset. Differentially expression gene (DEG) analysis was performed to both White and Black African American cancer and normal dataset as a reference. 143 and 1 DEG are found in White and Black African American race respectively. ENSG00000225937.1 (PCA3) is identified as the highest up-regulated gene expression in cancer in both White and Black African American race. The results of DEG analysis then become features to develop Artificial Intelligence (AI) classification system. AI model was developed using decision tree and random forest with GriDSearch parameters optimization and stratified 10-fold cross validation. Both Decision tree and random forest model yield 96% accuracy in training dataset and 93% and 91% accuracy in testing dataset for decision tree and random forest, respectively**

***Index Terms*— *Prostate cancer, Bioinformatics, Artificial Intelligence, Genomic. Machine Learning*.**

I. INTRODUCTION

Prostate cancer is one of deadly disease in a man. Based on cancer statistic 2024 estimated new cancer cases 299,010 and deaths 35,250 in United States [1]. However, the case of prostate cancer is also high in many others countries. Even though a clinical pathway that consist of diagnose, medical treatment and prognostic prediction of prostate cancer is already designed by world health organization (WHO) and health ministry in each country worldwide, the incidence and the mortality is still happened in high number of cases. Prostate cancer risk is 70% higher in black men compared to the white males [2]. Black men living in the United States having the highest risk at 4.2% [3]. Unfortunately, the standard blood test used to screen the prostate cancer, the prostate specific antigen (PSA) test is considered to be in accurate [4]. PSA is failing to identify 8 of every 10 men aged under 60 who later have prostate cancer diagnosed. Therefore, it is really urgent need to do a research to create an accurate prostate cancer test.

Genomic dataset has been used to predict, determine personal treatment and prognosis of a disease. Previously, bioinformatics study design was used to conduct genomic dataset analysis [5]–[8]. Currently, Deep learning and machine learning techniques has been integrated with bioinformatics research pipeline to explore the classification, treatment prediction and regression analysis [9], [10]. Nowadays, the knowledge and technology of genomics, bioinformatics and AI are ready to be used in research and project. However, the research that integrated Genomic, bioinformatics and AI in Health sector is still lack. Therefore, this study will focus utilizing the genomic dataset of prostate cancer to build a classification model utilized bioinformatics and AI study design.

The development of Genomic dataset, Bioinformatics software and tools, advance technology of Artificial Intelligence (AI) can be integrated to create a screening test for prostate cancer. Few studies start to explore a research utilizing genomic dataset, Bioinformatics and Artificial intelligence to develop an AI screening test for prostate cancer. Ramírez-Mena *et al* [11] utilizing 550 samples from TCGA to predict and identify prostate cancer tissue by gene expression with average sensitivity and specificity of 0.90 and 0.8 with AUC of 0.84. Moreover, Yousef et al [12] utilizing machine learning to classify prostate cancer or normal patients using gene - micro RNA (miRNA) pair dataset with the accuracy depends on gene – miRNA group. On

the group 10 with 1 miRNA and 122 genes achieved 0.95 AUC score. Another study by Kaplan and Ertunc [13] was also utilizing gene expression to diagnose prostate cancer that achieved 95.65% accuracy. However, none of those studies consider races as a unique variable in prostate cancer that affect black men incidence and death prostate cancer is higher than other races.

This research addressed racial issue in developing AI system by integrating several AI algorithm and bioinformatics approach to screening prostate cancer using genomic dataset. This study use gene expression and phenotype dataset to separated races and select features for develop AI system using Bioinformatics approach performed differentially expressed gene (DEG) analysis that yields up-regulated and down-regulated gene expression that can be used as features to develop AI system to classify cancer prostate or normal. Moreover, in the future this research design can be elaborated to predict the personal treatment and prognosis. In the future this study will be useful for the precision and personal medicine [14], [15].

II. METHOD

The overall research flowchart for this study is depicted in figure 1. It shows the summary of step by step in conducting the research.

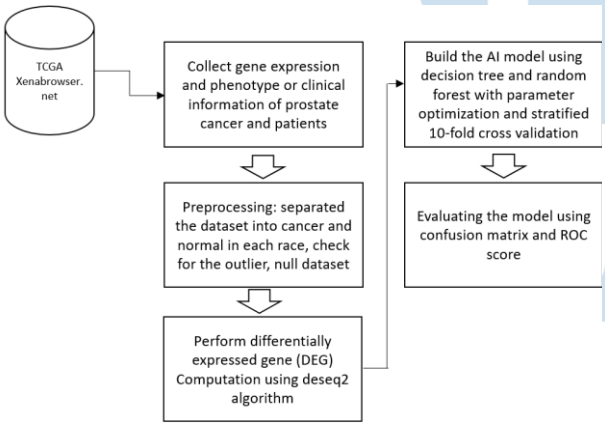

Figure 1. Research flowchart for early detection of prostate cancer using genomic, bioinformatics and AI study design

Data for this study were collected and downloaded from the TCGA project at https://xenabrowser.net/ on April 2024. The selected prostate study is by GDC TCGA using the code "PRAD". Two datasets were downloaded which were Phenotype and gene expression data. All downloaded datasets were saved locally. Data exploration, pre-processing, Bioinformatics computation and AI system development to classify prostate cancer and normal was performed using Python programming.

A. *Dataset Information*

The 551 datasets that were downloaded from the GDC TCGA Prostate Cancer (PRAD) on Xenabrowser.net at https://xenabrowser.net/ on April 2024. This research downloaded HTSeq – Counts gene expression RNAseq and Phenotype dataset. HTSeq – Counts gene expression RNAseq is a genome expression matrix dataset of patients. Moreover, Phenotype is clinical information related to the patients in HTSeq – Counts gene expression RNAseq dataset. Each patient has a unique TCGA ID.

B. *Data Exploration and preprocessing*

The dataset of HTSeq – Counts gene expression RNAseq were normalized. For data exploration and preprocessing, first, this study checks for missing value and outlier dataset. Second, matching TCGA ID between gene expression and phenotype dataset. Third, filtering the dataset based on cancer and normal. Moreover, filtering the data set in each cancer and normal dataset to a specific racial population. All the procedure in data exploration and preprocessing was conducted using Anaconda software and python programming.

C. *Bioinformatics computation Analysis*

This study utilized Bioinformatics methods for the features selection. Since the genome dataset consist of about 60,488 genes therefore this study will not use all the dataset. The differentially expression genes (DEG) dataset will be used for further analysis. DEG computation was performed using pydeseq2 package in python programming. Pydeseq2 is a package that implemented the mathematic procedural of deseq2 algorithm [16]. Commonly DEG calculation package is provided in R algorithm but we found the python version pydeseq2. Furthermore, DEG was performed for white and black race since another race like Asia is lack of the dataset. Top 3 up-regulated and down-regulated genes from DEG computation will be utilized as features to develop AI system.

D. *Develop AI system to screening prostate cancer in a spesific race*

Decision tree and Random forest algorithm was used to build AI system. Since we have imbalanced dataset between the class or label dataset therefore this study uses stratified k-fold cross validation that split the data such that the proportions between classes are the same in each fold as they are in the whole dataset. Stratified sampling is used to overcome the imbalanced issue without losing any inputs [17]. First, the dataset is separated into train and validation dataset. Train dataset will be used to develop the model and the validation dataset is used to test the model. Second, this study builds the model using decision tree and random forest algorithms since three-based algorithms often perform well on imbalanced datasets. Third, to optimize the accuracy, GridSearch parameter optimization was

performed. Lastly, using the best parameter from GridSearch this study build the AI model utilized stratified 10-fold cross validation. All the analysis in developed AI model was performed in Anaconda and python programming.

III. RESULTS AND DISCUSSIONS

A total of 551 patients' prostate cancer and normal was downloaded from xenabrowser.net. It consists of 498, 52, 1 cancer, normal and metastatic patients, respectively. Then, for cancer and normal patients we separated it again based on racial information. This study mapped expression dataset TCGA ID with phenotype dataset TCGA ID to separate the dataset per race. After mapped the TCGA ID this study found a total of 482 cancer patients that consist of 1, 10, 52, 406 patients for American Indian or Alaska native, Asian, Black or African and White race respectively. Moreover, there are 13 patients are not reported for the racial information. Furthermore, in normal population consist of 0, 0, 7, 44 for American Indian or Alaska native, Asian, Black or African and White race respectively and 1 patient is not reported for the racial information. It is shown that the number of genome expression dataset is still lack and still need further effort to make the data available especially for the normal dataset. For further analysis, to compute DEG, this study considers White and Black race. Furthermore, in developing the AI model this study only consider white race since the black race dataset is not proper for the AI model computation.

To compute DEG using pydeseq2, this study combined dataset with labelled information cancer and normal of each white and black race population. In white population the total dataset for DEG computation is 450 patients with 60,488 unique gene ID using ensemble ID. Then we deleted gene with the sum of the expression dataset from all the patients less or equal to 0. It yields only 57,412 genes. Then, baseMean, Log2FoldChange and P-adjusted value was performed to analyze differentially expression genes between cancer and normal population. The same procedure was also performed for the black race population. With the threshold P-adjusted value < 0.05 this study found that there are 143 and 1 genes are differentially expressed in white and black race respectively. DEG in cancer population has two types: (1) up-regulated and (2) down-regulated. Up regulated gene means the expression of a particular gene is higher than normal patients and it can become the potential cause of disease include cancer [18], [19]. In the other hand down regulated gene means the expression of a particular gene lower than the expression in normal that can cause a disease include cancer as well [20], [21] .It is shown that less results found in the black race compare to the white race. It can be caused by the lack of dataset in black race population. Interestingly, the gene that was found in black race ENSG00000225937.1 (PCA3) is also found in the white race with significantly up regulation expression. For further analysis, this study only focus on the white race population since the dataset is still proper for the analysis. Figure 2 is the distribution of the white race population of cancer and normal patients. This study plots all the raw value of input dataset of gene expression. It shows that the separation between patient's cancer and normal can be identified. However, the clustering is not well separated. Therefore, AI model for classification need to be developed.

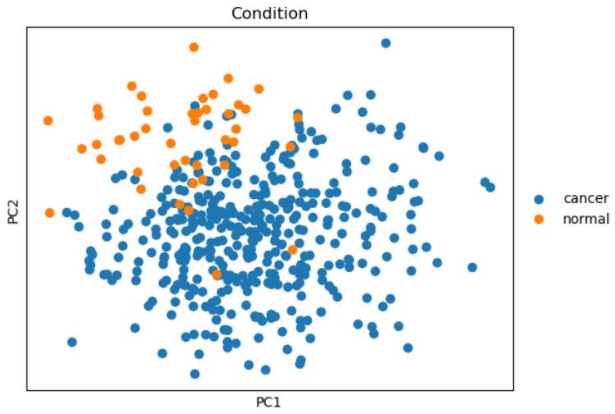

Figure 2. PCA analysis of dataset in white race cancer and normal population using gene expression value

Top 3 up-regulated gene and down-regulated gene resulted from deseq2 algorithm for DEG analysis as shown in table 1 was selected as the features to develop machine learning model for a classification prediction.

Table 1. Deseq2 computation results for white race

Ensembl_ID (Gene symbol)	baseMean	Log2FoldChange	P-adjusted value
ENSG00000225937.1 (PCA3)	12.3	0.6	7.5964E-13
ENSG00000242899.1 (RPL7P16)	11.2	0.4	1.0201E-06
ENSG00000166743.8 (ACSM1)	10.7	0.4	9.4623E-06
ENSG00000196878.11 (LAMB3)	10.6	-0.3	3.6607E-04
ENSG00000244509.3 (APOBEC3C)	10	-0.3	1.1646E-04
ENSG00000101443.16 (WFDC2)	10.2	-0.3	6.9417E-05

Table 1 shows top 3 up-regulated and down-regulated gene in white race prostate cancer compare to the normal population. Top 3 up-regulated gene are PCA3, RPL7P16 and ACSM1 while the top 3 down-regulated genes are LAMB3, APOBEC3C and WFDC2.

This study then evaluates whether those top-3 up-regulated and down-regulated gene can be used to classify the cancer and normal patients in prostate cancer white race population using machine learning analysis. This study use decision tree and random forest classifier. And label information is cancer or normal. This study first developed the model using

decision tree and random forest classifier with stratified 10-fold cross validation without parameters optimization. Stratified 10-fold cross validation was selected since the dataset is imbalanced as shown in figure 3.

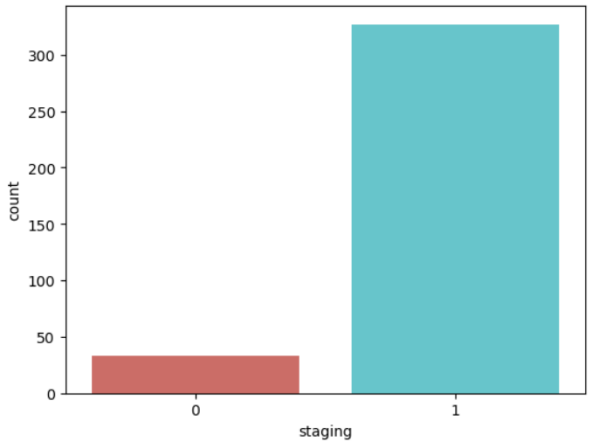

Figure 3. bar plot showing imbalanced dataset

Without GridSearch parameter optimization this study found the accuracy of the training model is 91.4% and 93.9% for decision tree and random forest classification respectively. Furthermore, GridSearch parameter optimization was performed to increase the accuracy of the classification method of decision tree and random forest. For decision tree, this study optimized four parameters criterion, max_depth, max_leaf_nodes and splitter. As the results the best value for each parameter can be seen in table 2. Moreover, for the random forest classification classifier Parameter optimization was also performed for criterion, max_features, n_estimators and min_sample_split and the best value is depicted in table 3.

Table 2. Best parameter optimization from GridSearch calculation for decision tree classifier

Parameter	Best value
Criterion	Entropy
Max_depth	2
Max_leaf_nodes	30
Splitter	Best

Table 3. Best parameter optimization from GridSearch calculation for random forest classifier

Parameter	Best value
Criterion	Entropy
Max_features	Log2
Min_samples_split	20
N_estimators	100

After parameter optimization was performed. The best parameters are used to build AI model using stratified 10-fold cross validation in decision tree and random forest algorithm. In decision tree, this study found there is a slightly increase accuracy score using parameter optimization. The comparison before and after parameter optimization can be seen in table 4.

Table 4. Comparison of evaluation results before and after parameter optimization for decision tree and random forest classifier

Model evaluation parameter	Before parameter optimization	After parameter optimization
Mean accuracy score of decision tree using stratified 10 fold cross validation	0.91	0.96
Mean accuracy score of random forest using stratified 10 fold cross validation	0.94	0.96

Table 4 shows that the training model for decision tree and random forest returned a high accuracy and there is no significant different between decision tree and random forest in building the training dataset AI model.

Test the model was performed using validation dataset. For decision tree for the test model evaluation is shown in table 5 and figure 4

Table 5. Evaluation results from the test dataset of decision tree classifier

Model evaluation type	Score
Accuracy	0.93
Precision	0.96
Recall	0.96
F1	0.96

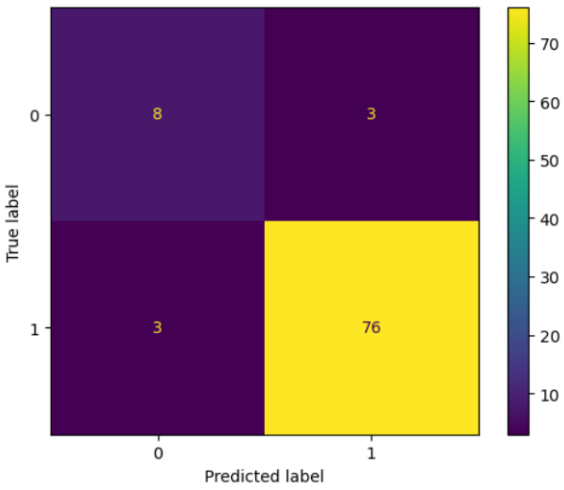

Figure 4. Decision tree confusion matrix from test dataset

Moreover, to evaluate the test result in decision tree scenario, AUC analysis was also performed as shown in figure 5.

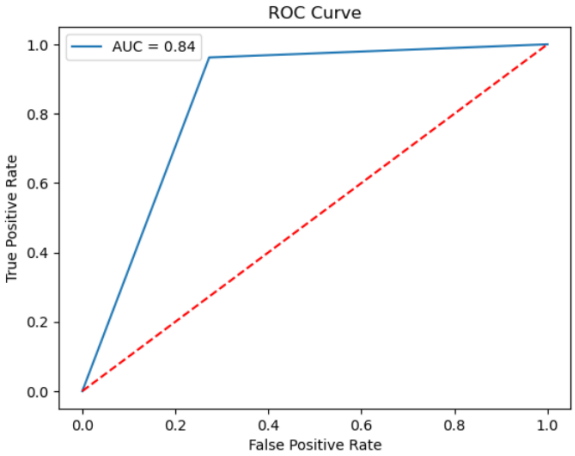

Figure 5. Decision tree ROC curve from test dataset

Figure 5 shows that the AUC score achieved 0.84. it indicates that the algorithm performs pretty well in separating the cancer and normal dataset.

This study was also test the model using random forest classifier and the model evaluation is depicted in table 6 and figure 6

Table 6. Evaluation results from the test dataset of random forest classifier

Model evaluation type	Score
Accuracy	0.91
Precision	0.95
Recall	0.95
F1	0.95

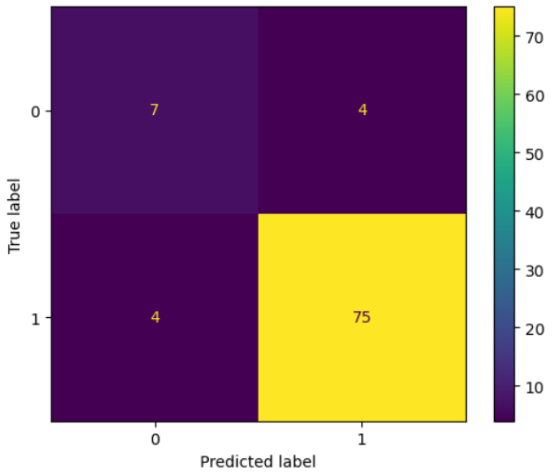

Figure 6. Random forest confusion matrix from test dataset

AUC analysis was also performed for random forest evaluation model for test dataset as shown in figure 7.

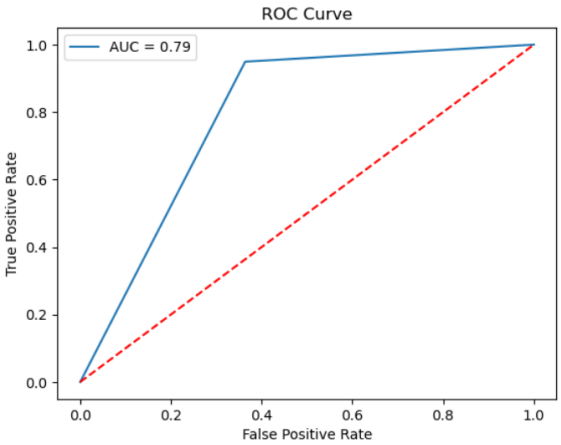

Figure 7. random forest ROC curve from test dataset

Figure 7 shows that the AUC score achieved 0.79. it indicates that score is slightly decreasing compare to the decision tree result.

It is shown that with the new dataset that come from the validation dataset the system still can predict in high accuracy 0.93 and 0.91 using decision tree and random forest algorithm classification classifier. It is shown that our developed model has a very strong dataset and algorithm that can be used as a tool to predict cancer or normal patients in prostate cancer. Through this study as well it is shown that gene expression dataset has a strong or powerful pattern to be used as a features for AI to diagnose a disease

IV. CONCLUSION

This research utilized prostate cancer genomic dataset to classify cancer or normal using bioinformatics and AI approach. The results show that with top 3 up and down regulated genes in specific race are able to screening the prostate cancer from normal with higher accuracy score in training, testing and validation dataset.

Through this study also we can understand that combining bioinformatics approach for feature filtering and selection is a powerful method in selecting the features for AI model development to predict a particular disease.

ACKNOWLEDGEMENT

DA performed major parts of the research and prepared the manuscript. AAP, MVO provide a review on the study methodology and manuscript writing. All authors read and approved the final manuscript. The authors are thankful for Department of Informatics, Multimedia Nusantara University, Serpong Banten, Indonesia, for the support in facility and financial under

internal grant number 0020-RD-LPPM-UMN/P-INT/VI/2024 for the whole research process.

REFERENCES

[1] R. L. Siegel Mph, A. N. Giaquinto, | Ahmedin, J. Dvm, and R. L. Siegel, "Cancer statistics, 2024," *CA. Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, Jan. 2024, doi: 10.3322/CAAC.21820.

[2] "Press Releases." https://pressroom.cancer.org/ProstateCancerMortality (accessed Aug. 08, 2024).

[3] M. A. Jain, S. W. Leslie, and A. Sapra, "Prostate Cancer Screening," *StatPearls*, Oct. 2023, Accessed: Jul. 20, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK556081/

[4] S. Gottlieb, "Study shows poor reliability of prostate cancer test," *BMJ Br. Med. J.*, vol. 327, no. 7409, p. 249, Aug. 2003, Accessed: Jul. 20, 2024. [Online]. Available: /pmc/articles/PMC1126651/

[5] A. F. Liko, E. Ciputra, N. A. Sanjaya, P. C. Thenaka, and D. Agustriawan, "DNA methylation profiling reveals new potential subtype-specific gene markers for early-stage renal cell carcinoma in Caucasian population," *Quant. Biol.*, vol. 10, no. 1, pp. 79–93, Mar. 2022, doi: 10.15302/J-QB-021-0279.

[6] K. N. Ramanto, K. J. Widianto, S. S. H. Wibowo, and D. Agustriawan, "The regulation of microRNA in each of cancer stage from two different ethnicities as potential biomarker for breast cancer," *Comput. Biol. Chem.*, vol. 93, Aug. 2021, doi: 10.1016/J.COMPBIOLCHEM.2021.107497.

[7] J. Ivan, G. Patricia, and D. Agustriawan, "In silico study of cancer stage-specific DNA methylation pattern in White breast cancer patients based on TCGA dataset," *Comput. Biol. Chem.*, vol. 92, Jun. 2021, doi: 10.1016/J.COMPBIOLCHEM.2021.107498.

[8] M. Gustiananda *et al.*, "Immunoinformatics Identification of the Conserved and Cross-Reactive T-Cell Epitopes of SARS-CoV-2 with Human Common Cold Coronaviruses, SARS-CoV, MERS-CoV and Live Attenuated Vaccines Presented by HLA Alleles of Indonesian Population," *Viruses*, vol. 14, no. 11, Nov. 2022, doi: 10.3390/V14112328.

[9] Z. Zhang *et al.*, "Deep learning in omics: a survey and guideline," *Brief. Funct. Genomics*, vol. 18, no. 1, pp. 41–57, 2019, doi: 10.1093/bfgp/ely030.

[10] Q. Wu *et al.*, "Deep Learning Methods for Predicting Disease Status Using Genomic Data," *J. Biom. Biostat.*, vol. 9, no. 5, 2018, Accessed: Jul. 29, 2024. [Online]. Available: /pmc/articles/PMC6530791/

[11] A. Ramírez-Mena, E. Andrés-León, M. J. Alvarez-Cubero, A. Anguita-Ruiz, L. J. Martinez-Gonzalez, and J. Alcala-Fdez, "Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression," *Comput. Methods Programs Biomed.*, vol. 240, Oct. 2023, doi: 10.1016/J.CMPB.2023.107719.

[12] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, "Mircornet: Machine learning-based integration of mirna and mrna expression profiles, combined with feature grouping and ranking," *PeerJ*, vol. 9, p. e11458, May 2021, doi: 10.7717/PEERJ.11458/TABLE-12.

[13] K. Kaplan and H. M. Ertunc, "Diagnosis of prostate cancer using gene expression data," *26th IEEE Signal Process. Commun. Appl. Conf. SIU 2018*, pp. 1–4, Jul. 2018, doi: 10.1109/SIU.2018.8404250.

[14] O. Strianese *et al.*, "Precision and Personalized Medicine: How Genomic Approach Improves the Management of Cardiovascular and Neurodegenerative Disease," *Genes (Basel).*, vol. 11, no. 7, pp. 1–24, Jul. 2020, doi: 10.3390/GENES11070747.

[15] S. J. Maceachern and N. D. Forkert, "Machine learning for precision medicine," *Genome*, vol. 64, no. 4, pp. 416–425, 2021, doi: 10.1139/GEN-2020-0131/ASSET/IMAGES/LARGE/GEN-2020-0131F1.JPEG.

[16] D. Rosati *et al.*, "Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1154–1168, Dec. 2024, doi: 10.1016/J.CSBJ.2024.02.018.

[17] X. Zhao, J. Liang, and C. Dang, "A stratified sampling based clustering algorithm for large-scale data," *Knowledge-Based Syst.*, vol. 163, pp. 416–428, Jan. 2019, doi: 10.1016/J.KNOSYS.2018.09.007.

[18] E. Scott *et al.*, "Upregulation of GALNT7 in prostate cancer modifies O-glycosylation and promotes tumour growth," *Oncogene*, vol. 42, no. 12, p. 926, Mar. 2023, doi: 10.1038/S41388-023-02604-X.

[19] E. D. Arisan *et al.*, "Upregulated Wnt-11 and miR-21 Expression Trigger Epithelial Mesenchymal Transition in Aggressive Prostate Cancer Cells," *Biol. 2020, Vol. 9, Page 52*, vol. 9, no. 3, p. 52, Mar. 2020, doi: 10.3390/BIOLOGY9030052.

[20] N. C. Whitlock *et al.*, "MEIS1 down-regulation by MYC mediates prostate cancer development through elevated HOXB13 expression and AR activity," *Oncogene*, vol. 39, no. 34, p. 5663, Aug. 2020, doi: 10.1038/S41388-020-01389-7.

[21] V. Nodouzi, M. Nowroozi, M. Hashemi, G. Javadi, and R. Mahdian, "Concurrent Down-Regulation of PTEN and NKX3.1 Expression in Iranian Patients with Prostate Cancer," *Int. braz j urol*, vol. 41, no. 5, pp. 898–905, 2015, doi: 10.1590/S1677-5538.IBJU.2014.0036