

Original Paper

Adithama Mulia¹, David Agustriawan^{1*}, Marlinda Vasty Overbeek¹, Moeljono Widjaja¹, Vincent Kurniawan¹, Jheno Syechlo¹, Muhammad Imran Ahmad², Srinivasulu Yerukala Sathipati³

¹Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Indonesia.

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

³Marshfield Clinic Research Institute, Marshfield Clinic Research Institute, Marshfield Wisconsin 54449

*Corresponding Author. Email: david.agustriawan@umn.ac.id. Phone: +62 877 8153-5936.

Artificial Intelligence Design for Racial-based Prostate Cancer Stage Classification with Multi-Layer Perceptron: Feature Selection Optimization Approach

Abstract

Background: Prostate cancer progression exhibits significant variability influenced by biological and racial factors. DNA methylation profiling has shown potential in early cancer detection, but its integration with machine learning across racially diverse populations remains limited.

Objective: This study aims to develop a race-aware framework using DNA methylation data and a Multi-Layer Perceptron (MLP) model to classify prostate cancer stages into early (I-II) and late (III-IV) stages.

Methods: Methylation and phenotype data from the TCGA-PRAD dataset were processed using Differentially Methylated Positions (DMP) analysis to identify CpG sites correlated with cancer stages. These features were further refined through Recursive Feature Elimination (RFE) and used to train MLP models.

Results: The best-performing model achieved ~95% accuracy and up to 99% AUC on the majority race (White) training data using 70 selected features. However, performance declined sharply on minority race groups, revealing the effects of sample imbalance and race-specific methylation patterns.

Conclusions: MLP models based on DNA methylation data show strong potential for accurate and cost-efficient early-stage prostate cancer detection. However, racially diverse datasets are crucial to ensure fairness and generalizability in predictive performance.

Keywords: Prostate Cancer, Multi-Layer Perceptron, Feature Selection, DNA Methylation, Differentially Methylated Positions, Race-aware

Introduction

Prostate cancer is the most common organ-specific cancer and the second leading cause of cancer-related death among men in the United States [1,2], with over 191,930 new cases reported in 2019 and an estimated 33,310 deaths in 2020 [3]. Risk factors such as age, family history, and lifestyle contribute to its high incidence, especially among older men [4]. Biologically, it is a heterogeneous disease, exhibiting diverse subtypes and gene expression patterns that affect prognosis and treatment [5–7]. In addition, prostate cancer has several biological variants that exhibit different levels of severity and response to treatment. These variations can influence the course of the disease, ranging from slow-growing forms to more aggressive and fast-spreading. This biological complexity presents a challenge for achieving accurate and consistent diagnosis across diverse patient populations.

Classification of prostate cancer typically depends on the Gleason score and androgen receptor (AR) status, which inform tumor aggressiveness and therapeutic response [8]. Genetic mutations including TMPRSS2-ERG, PTEN, and SPOP have also been implicated in disease progression [9–11]. Advances in sequencing and molecular profiling have enabled the discovery of new biomarkers for targeted treatment [12], yet many classification models still fail to capture the nuanced biological variation across cancer stages.

Next-Generation Sequencing (NGS) have seen rapid technological advancements along the years and has become central to exploring epigenetic regulation in cancer [13]. DNA methylation, an epigenetic mechanism that modifies gene expression without altering the sequence, is key to identifying biomarkers and understanding cancer development. NGS enables high-throughput, genome-wide methylation profiling, facilitating the discovery of fine-grained methylation patterns and their association with specific mutations [14–16]. However, the vast complexity of methylation data demands computational approaches that can extract clinically actionable insights, which remains a key challenge.

Methylation data analysis especially DMPs address the high dimensionality of methylation profiles by identifying individual CpG sites with significant methylation differences between groups [17]. DMP analysis offers single-site resolution and is better suited for classification tasks compared to region-based methods like DMRs, especially when individual CpG sites are used as input features [18]. Statistical methods such as t-tests and linear models help quantify these differences [19]. Despite their advantages, DMPs are rarely used in conjunction with machine learning models like MLP for stage-specific cancer classification, highlighting a gap that this research aims to address.

Multi-Layer Perceptron (MLP) is a neural network model widely used for pattern recognition and classification in fields like bioinformatics, particularly cancer diagnostics [20–23]. It can effectively manage high-dimensional data such as DNA methylation, outperforming many traditional statistical approaches. MLPs have

been shown to identify cancer-specific biomarkers and classify cancer stages with high accuracy [24,25]. Beyond that, MLP also support personalized therapy by improving diagnostic consistency. However, no prior studies have combined DMP-based feature selection with MLP for stage-specific cancer classification, making this approach a novel contribution to the field. This framework provides accurate racial-based classification of prostate cancer for diagnosed individuals, offering an effective tool for rapid initial screening and preoperative assessment.

Although machine learning has significantly advanced cancer classification, current genomic-based models often neglect racial disparities, particularly in DNA methylation patterns. A study by Abdollahi et al. [26] introduces a radiomic-based framework for prostate cancer evaluation, utilizing machine learning techniques and magnetic resonance imaging (MRI) features to predict treatment response, Gleason score (GS), and cancer stage. The study involves 33 prostate cancer patients who went through pre- and post-intensity-modulated radiation therapy (IMRT) T2-weighted (T2W) and apparent diffusion coefficient (ADC) MRI scans. Radiomic features were extracted from both image types, and univariate analysis combined with paired t-tests identified significant features associated with treatment response. Feature selection and classification were performed using tenfold cross-validation across various image sets, leading to post-T2W radiomic models achieving the highest predictive performance with an AUC of 0.632, followed by pre-ADC (AUC 0.626) and pre-T2W (AUC 0.61). Additionally, T2W-based models yielded a mean AUC of 0.739 for GS prediction, while ADC-based models performed better in stage prediction with a mean AUC of 0.675. Another study by Hartenstein et al. [27] evaluates three convolutional neural networks (CNN) trained to detect lymph node infiltration (LNI) by prostate cancer using contrast-enhanced CT images. The models were tested against expert radiologists and achieved comparable performance. The best-performing CNN, trained on a status-balanced dataset, reached an AUC of 0.90, primarily by learning anatomical context. In contrast, a location-balanced CNN and a segmentation-masked (xMask) CNN reached lower AUCs of 0.858 and 0.677, respectively. Random forest classifiers using only nodal volume and location performed well on status-balanced data (AUC 0.90), but poorly on location-balanced data (AUC 0.677), reinforcing that CNNs leveraged anatomical features in classification. A similar study by Eissa et al. [28] utilized DNA methylation profiles to investigate the impact of feature selection in reducing the dimensionality of methylation data through metaheuristic techniques. The study was structured in two stages: the first focused on feature selection, and the second on developing a Deep Neural Network (DNN) model to classify samples based on malignancy status and cancer type. The proposed method achieved competitive results compared to existing approaches, with strong performance in terms of recall, precision, and accuracy, along with excellent receiver operating characteristic area under the curve (ROC AUC) values ranging from 0.85 to 0.89. While previous studies have achieved promising results using artificial neural network (NN) models, most have focused on imaging data, and the use of DNA methylation profiles as features remains relatively uncommon. Among the limited research utilizing methylation data, few have attempted to classify malignant cancer stages, particularly in

distinguishing early from late-stage disease. Additionally, many models do not prioritize minimizing feature sets, which may reduce model interpretability and increase the risk of overfitting. Racial disparities in DNA methylation patterns are also often overlooked, with limited efforts to develop race-specific models despite known imbalances in publicly available datasets. To address these gaps, we utilize DNA methylation data and apply a combination of DMP and RFE to identify a minimal, high-impact set of features. We further validate the selected biomarkers using the Catalogue Of Somatic Mutations In Cancer (COSMIC) and cBioPortal.org to ensure biological relevance. Our study introduces a race-aware MLP model focused on prostate cancer stage classification specifically distinguishing early from late stages demonstrating the importance of both racial context and minimal feature sets in developing more accurate and equitable diagnostic tools.

Methods

Study Design

This study implements data collection, preprocessing, feature selection, and MLP modeling and evaluation as seen in Figure 1. These methods are conducted using Python (version 3.12.3; Python Software Foundation) programming language and the necessary libraries using Visual Studio Code editor (version 1.95.3; Microsoft Corp).

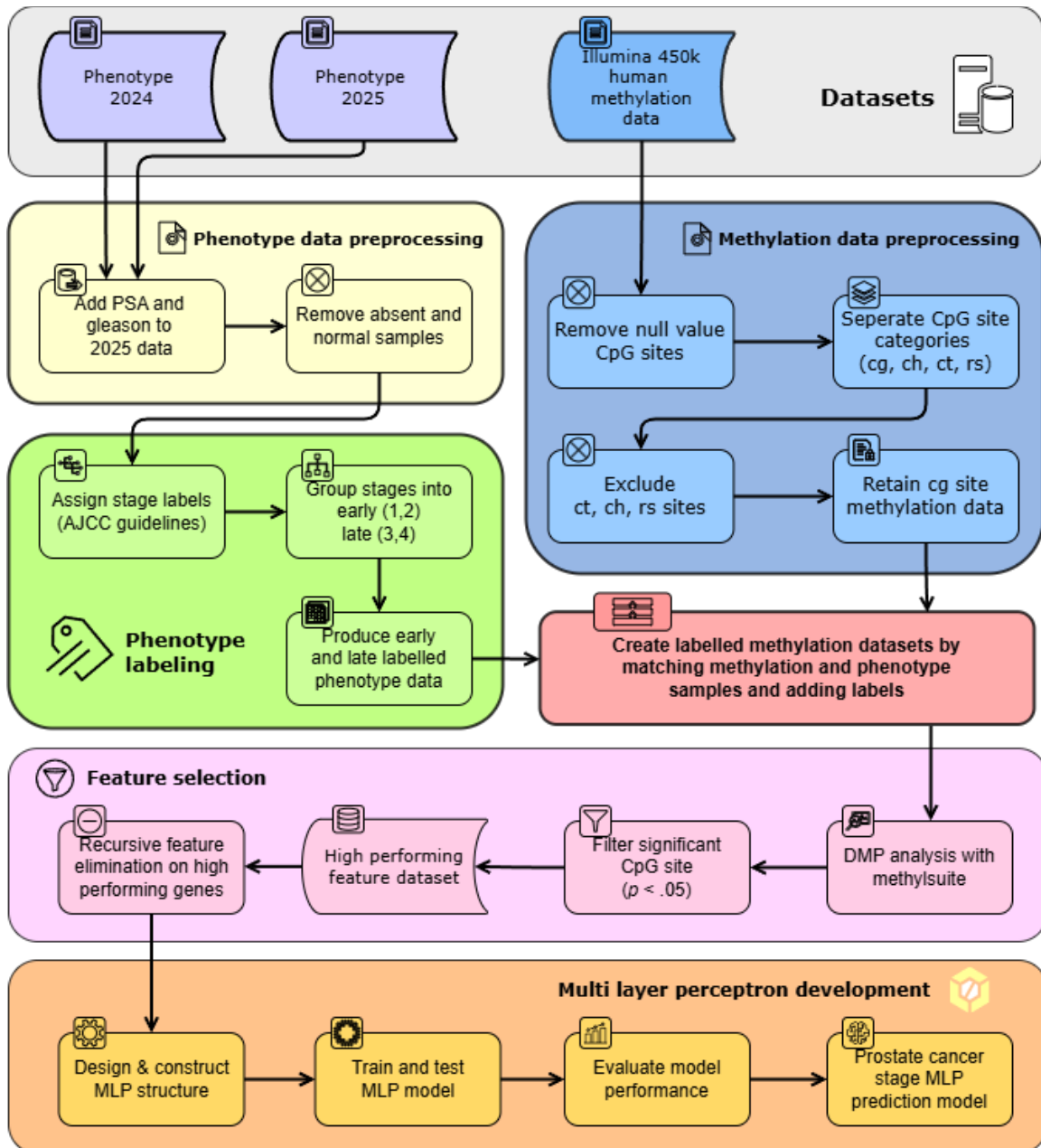


Figure 1. Datasets are processed differently, phenotypes are combined to create a complete version appropriate for labelling process. Methylation data is separated by their unique CpG sites (cg, ch, ct, rs) however, only cg is used in this research as the remaining CpG sites lacks information on methylation values. Samples in phenotype dataset is labelled according to AJCC Cancer Staging Guidelines 7th edition and simplified into 2 groups, “early” (stage 1 & 2) and “late” (stage 3 & 4). Both datasets are then matched on their samples and a label column is added to methylation dataset. Feature selection utilizes the labelled methylation dataset to analyze each CpG site’s importance through DMPs and RFE process. Remaining CpG sites are used as features in MLP model development to train and predict prostate cancer stages.

Data Gathering and Preparation

The data used in this study were obtained from the UCSC Xena Browser, specifically from the Genomic Data Commons (GDC) The Cancer Genome Atlas - Prostate Adenocarcinoma (TCGA-PRAD) cohort [29]. This includes Illumina HumanMethylation450 BeadChip data and the corresponding phenotype information, both downloaded in February 2025. However the latest phenotype presents missing fields essential to the labelling process. To supplement missing fields in the latest phenotype release, particularly those essential for the labeling process, thus an earlier version of the phenotype dataset retrieved in early 2024 was also incorporated. A completed version of the phenotype dataset is available on **Multimedia Appendix 2**.

The illumina 450k DNA methylation dataset consists of approximately 450k CpG sites with beta values ranging from 0~1 indicating methylation status of hypermethylated or hypomethylated. However, in the dataset there are a number of CpG sites for which all values are null for all samples and partially null in some samples, indicating that no methylation data is available at all for that site. In this case, not only are the null cells ignored, but all columns of completely empty CpG sites are excluded from the feature selection process and partially empty CpG sites are replaced with 0 to indicate that the site is not methylated, ensuring that only features with valid information are included in the subsequent analysis.

The DNA methylation dataset consists of four types of CpG sites, namely cg, ct, ch, and rs, each representing a specific location on the DNA where methylation occurs. Among the four types, cg-type CpG sites are the most dominant, accounting for more than 90% of the data. In contrast, CpG site types ct, ch, and rs have high levels of missing values, with ch containing more than 60% null values and rs being entirely null. Due to the high proportion of missing values in these three types, and also to maintain data quality and consistency, this study only used features derived from CpG sites of type cg, while features of type ct, ch, and rs were excluded from the analysis. After obtaining the cg dataset, the samples were separated based on each race in the phenotype data. The race with the highest number was used as the data for training and testing. Each race dataset presents partially missing values in the cg data, to handle this, the mean imputation method is used to compensate for the missing values [30,31]. The remaining racial samples were used as external validation of the possibly racially based characteristics of the methylation data.

Sample Labelling

This study begins by labeling each sample in the dataset based on cancer stage. Labeling is done using the prostate cancer labeling manual from the American Joint Committee on Cancer (AJCC) Cancer Staging Manual 7th edition page 473 can be seen in Figure 2. There are 5 stages of prostate cancer staging, in this labeling process it is shortened by combining stages IIA and IIB into one class, namely stage II, reducing the total existing classification to 4 and labeling is carried out based on these 4 stages which are then further shortened accordingly, stages 1 and 2 are

classified as early because of the nature of the stage where the tumor has not spread to areas other than the prostate as stages 3 and 4 are classified as late because at that stage the tumor has spread to lymph nodes or metastasized [32]. By referring to several criteria that distinguish cancer stages in the AJCC Cancer Staging Manual; T (presence of tumor), N (spread to lymph nodes), M (metastasis), PSA and Gleason. However, since the PSA (Prostate Specific Antigen) and Gleason criteria were missing from the Xenabrowser phenotype dataset in 2025, this study compensated the missing criteria from the Xenabrowser phenotype dataset in 2024.

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

Figure 2. AJCC Instruction to Prostate Cancer Stage Groups [32]. AJCC guidelines stratifies patients into distinct groups (I, IIA, IIB, III, IV) based on key diagnostic parameters: T (tumor size and extent), N (regional lymph node involvement), M (distant metastasis), PSA (Prostate-Specific Antigen) levels, and Gleason score. These parameters collectively inform the cancer's aggressiveness and spread. However, this research follows a common rationale for consolidating Stage IIA and IIB into a singular group "Stage II" is based on the shared characteristic: the tumor remains primarily confined within the prostate (T1 or T2) without any evidence of spread to

regional lymph nodes (N0) or distant sites (M0). Despite minimal differences in tumor volume or specific T2 sub-classifications between IIA and IIB, their localized nature justifies a combined classification for simplified communication and initial treatment planning.

The distinction between "early" stages (Stage I and II) and "late" stages (Stage III and IV) is critically important. Early stages, consisting of Stage I (very early, often microscopic, localized within the prostate) and Stage II (cancer confined within the prostate capsule), are characterized by localized diseases that have not spread to lymph nodes or distant organs. This localization makes these stages highly manageable to treatments focused on the prostate, such as surgery or radiation, typically leading to a favorable prognosis [33]. In contrast, late stages, including Stage III (cancer extending just beyond the prostate capsule) and Stage IV (cancer spreading to regional lymph nodes or distant organs), represent more advanced, often systemic disease. At these stages, the cancer's spread makes complete eradication through local treatments less likely, often involves systemic therapies to manage the widespread disease and generally resulting in a less favorable prognosis.

Feature Selection

Feature selection begins with a DMPs analysis of all genes against the label to find methylation values that are outliers against the label. The DMPs analysis was run with the MethySuite python package using the function `methySuite.diff_meth_pos.diff_meth_pos` which accepts sample ID, beta value, and label. The function returns the DMPs analysis results consisting of columns; cg pos, P-val. Then from the analysis results, a cut-off limit is given at $P < .05$. The number of cg obtained from the analysis is used as a feature for the MLP model. In order to minimize the number of features, the RFE algorithm is further implemented by using SVC algorithm as classifier. The RFE process produces several subsets of data that are used to evaluate the minimum number of features, where too few features cause the model to have difficulty in distinguishing between classes. To validate the biological significance of selected CpG sites, genes related to the CpG site were tracked using Infinium HumanMethylation450 v1.2 BeadChip Product Files [34].

To validate the results of feature selection, gene mutation checks were performed on all selected CpG sites. Gene mutation checking starts by taking the Illumina ID of the CpG site (example: cg00000029) and matching the Illumina ID with the gene symbols available from the Infinium HumanMethylation450 v1.2 BeadChip Product Files. After obtaining a list of gene symbols, checking gene mutations can be done by querying to an external dataset of cancer genes known to have mutations in prostate cancer cases by COSMIC, matching gene symbols in the dataset and calculating the percentage of individual gene mutations from the number of samples that have been tested with the number of samples recorded mutations [35]. Further validation can be done on the cbiportal.org website by querying gene symbols in the Prostate Adenocarcinoma dataset (TCGA, PanCancer Atlas) to obtain more detailed

information about gene mutations per prostate cancer stage in the sample dataset used in this study (TCGA) [36,37].

MLP Development

Construction of the MLP model starts with determining the scheme of the MLP to be used. Generally, one to two hidden layers are used because numerical/tabular data tends to have a simpler structure than image or text data. The number of neurons in each layer is usually determined heuristically, for example by taking the average of the number of input and output neurons, or using a decreasing pattern such as 32-16-8. Activation functions such as ReLU are commonly used in the hidden layer to improve the ability to represent non-linear data, while sigmoid or softmax activation functions are used in the output layer according to the type of classification, in this case it is a binary classification so sigmoid activation functions are used for all output layers. The final architecture selection is generally determined through experimentation and validation to avoid overfitting and ensure optimal model performance.

After determining the MLP model scheme, model training begins by separating the dataset into training and testing data. The training and testing data is obtained from separating the initial dataset. The cross-validation method is implemented to the training process to prevent potential overfitting or bias to the training data thus expanding the generalizability of the model. Model training is performed incrementally in epochs until the process is automatically stopped by an early stopping mechanism. This mechanism stops training if there is no improvement in the model's performance on the validation data after a certain number of epochs, thus preventing overfitting. During the training process, the model is evaluated using Adam's optimizer. All training scenarios use a consistent loss function, binary cross entropy, in accordance with the nature of the binary classification problem at hand.

Performance Evaluation

Model evaluation is performed using the `classification_report` function, focusing on harmonization between the f1-score, accuracy, precision, and recall metrics. Model performance is evaluated based on results on training and testing data to ensure reliability without overfitting or underfitting. In addition, the Area Under the Curve (AUC) value is also used as an additional metric to assess the model's ability to distinguish between classes as a whole.

Results

Prepared Data

The data for this study consisted of two correlated secondary datasets. The datasets were obtained through the public databases Xenabrowser GDC TCGA-PRAD DNA methylation - Illumina Human Methylation 450 and Xenabrowser GDC TCGA-PRAD phenotype in 2024 and 2025. The DNA methylation - Illumina Human Methylation

450 dataset contains 553 samples with TCGA barcodes and 486,428 identifiers or CpG sites with beta values (0~1) as shown in Table 1.

Table 1. DNA Methylation dataset matrix example.

	TCGA-G9- A9S7-01A	TCGA-EJ- 7792-11A	TCGA-EJ- 7792-01A	...	TCGA-J4- A67L-01A
cg00000029	Null	0.1888	0.2635		0.1521
cg00000108	0.9691	0.9670	0.9685		0.9726
...					
	Methylation	Methylation	Methylation		Methylation
cg-N	value	value	value	...	value
	cg-N	cg-N	cg-N		cg-N

^aThe values in the matrix are each sample's CpG site methylation value represented in beta values ranging from 0 to 1.

The GDC TCGA-PRAD Phenotype 2024 and 2025 datasets have 572 samples in the form of TCGA barcodes and 88 identifiers in the form of categorical clinical data as shown in **Multimedia Appendix 2**.

The initial DNA methylation dataset consists of a total of 486,428 CpG site features and a total of 893 samples and the phenotype data only has 572 samples. From both datasets, this study only takes samples that are in both datasets, so from the methylation data which has 553 only 415 samples are taken in as this research only uses white population from both phenotype data by equalizing the TCGA barcode of both datasets. After identifying the null values, it was found that 64,728 CpG sites had null values in all samples, which means there is no methylation value at all at the site. Such features were considered to have no analytical value and were excluded entirely from the dataset.

After the removal of completely empty CpG sites, the number of features was reduced to 421,699 features, or about 87% of the original data and then the samples were separated by race. White race is the dominant race in the dataset, covering up to 415 out of 572 total samples, and was therefore used as the basis for the MLP model construction process. The remaining samples with other races were used for validation of the hypothesized race-based characteristics of the dataset. The races used as validation data are Asian and black or African American with 12 samples in Asian race and 60 samples in black race. There were races in the phenotypes that were not covered by this study due to 'not reported' race values and insufficient number of labels to predict. This process ensured that only features with valid and analyzable methylation information were used in the DMPs and RFE analysis stages.

Labelling

Based on the labeling performed using the AJCC Cancer Staging Manual guidelines, each sample in the dataset was successfully classified into the appropriate prostate

cancer stage. Some samples had missing values in one or two of the main criteria (T, N, or M), but no samples had missing values in all five criteria at once. To handle this, missing values in the T, N, and M categories were treated as T0, N0, and M0, which were assumed to be conditions without significant spread or growth. Meanwhile, the PSA and Gleason columns have no missing values and can be fully used in labeling. From a total of 415 white race samples, the label distribution was obtained as represented in Figure 3.

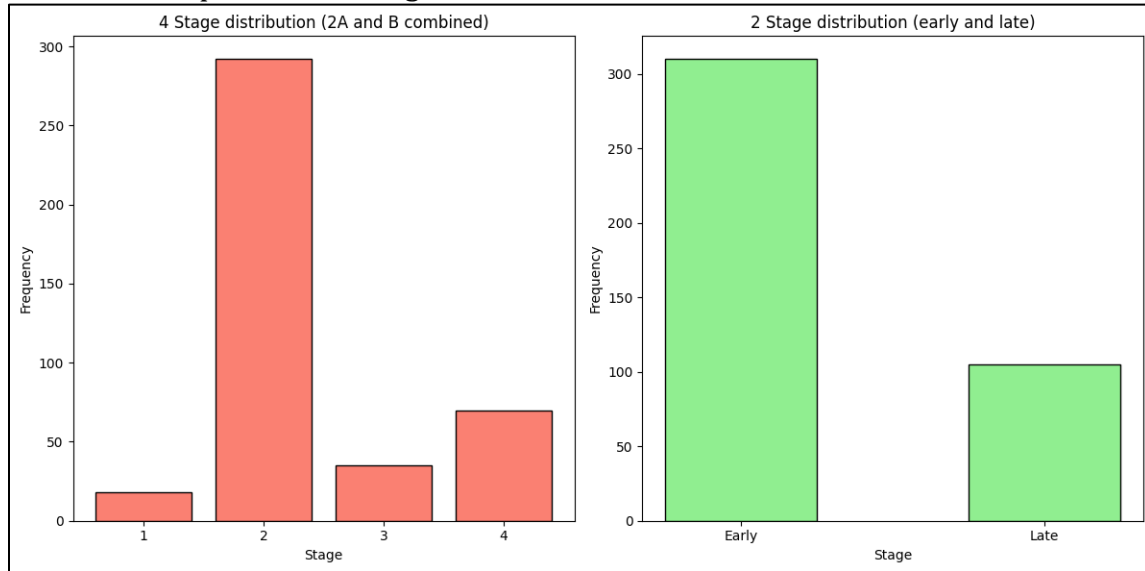


Figure 1. Distributions of sample stage groups. Stage IIA and IIB were combined to form a dominant Stage II group, comprising 292 samples. In contrast, the other stage groups contain significantly fewer samples: Stage I with 18 samples, Stage III with 35 samples, and Stage IV with 70 samples. For simplification, these stages were further grouped into two broader categories: “early” (Stages I and II) with a total of 310 samples, and “late” (Stages III and IV) with 105 samples.

This distribution shows a predominance of samples at advanced stages (Stage 2), reflecting possible population bias or the availability of more data for advanced cases. The labeling process was done manually and adjusted to the AJCC classification scheme to ensure accuracy and consistency between samples. All label data was then used for further analysis, including classification or clustering of prostate cancer stages.

Feature Selection

Feature selection is performed in two stages, namely DMPs analysis using methylsuite and advanced feature reduction using RFE algorithm. The goal of this process is to obtain a minimum subset of CpG site features that are most relevant to prostate cancer staging.

The implementation of DMPs using methylsuite starts by excluding CpG sites that have been recorded as insignificant or no correlation to oncology from journals or public databases which eliminates ~200,000 CpG sites. The remaining ~280,000 features were then subjected to DMPs analysis using a p-value cutoff of $< .05$ to

evaluate their influence on the label. DMPs analysis with a cutoff of $P < .05$ resulted in 6,501 CpG sites that were significant to the label. The cutoff of $P < .05$ was chosen as the default value because it is considered to provide a balance between selectivity and the number of features that can still be processed efficiently.

Then, the RFE algorithm was used to evaluate and select the most informative features from the 6,501 CpG sites. RFE is applied by gradually generating several feature subsets, namely by reducing the number of features to 100, 90, 80, 70, and 60. Each of these feature subsets is used to form a different dataset for each race that will be used for training and testing classification models. Each dataset has its own characteristics depending on the number of features and estimators used in the RFE process. The final feature selection considers the balance between model complexity and classification performance, with the aim of avoiding the risk of overfitting or underfitting.

Biological relevance of features to labels is validated by checking several online datasets. First searching for reference genes from 100 selected CpG sites. The reference gene is obtained from the Infinium HumanMethylation450 v1.2 BeadChip manifest file by matching the CpG site to the gene symbol. The results of matching CpG sites with gene symbols in manifest files obtained 58 symbols from sites that have reference genes. Then 58 gene symbols were matched to the COSMIC Cancer Browser tool by selecting prostate cancer tissue and exporting all data about genes that are often found in prostate cancer and recording the number of samples and the number of mutations that occur. Genes with a mutation rate above 5% data is shown in Table 2.

Table 2. Genes with a percentage of mutations above 5% in prostate cancer

Gene	Mutated samples	Samples tested	Mutation Rate
MACROD2	616	2154	28.60%
MAGI2	539	2212	24.37%
FAM155A	492	2154	22.84%
DLGAP2	477	2154	22.14%
FMN2	321	2154	14.90%
NTM	305	2154	14.16%
L3MBTL4	291	2154	13.51%
SHISA9	276	2154	12.81%
PCDHGA4	218	2154	10.12%
FBXL7	189	2154	8.77%
GABRB3	165	2154	7.66%
PCDH17	127	2154	5.90%

^aBased on the COSMIC database, several genes showed low mutation frequencies in prostate cancer samples. MACROD2 led the way with a mutation rate of 28.6%, followed by MAGI2 (24.37%), FAM155A (22.84%), and DLGAP2 (22.14%). These 4

genes alone are mutated in more than 1 in 5 cases, indicating a potential role in the pathogenesis of prostate cancer.

Based on the COSMIC database, the selected genes shows low mutation frequencies in prostate cancer samples. MACROD2 leads with a mutation rate of 28.6%, followed by MAGI2 (24.37%), FAM155A (22.84%), and DLGAP2 (22.14%). These four genes alone are mutated in more than one out of five cases, suggesting a potential role in the pathogenesis of prostate cancer.

While classical tumor suppressors such as TP53 and PTEN are not included in this list, the mutation rates observed here may reflect less well-known but recurrent alterations in prostate cancer stages. Notably, genes such as FMN2, NTM, and L3MBTL4 also show relatively comparable mutation rates, raising questions about their functional significance. However, COSMIC does not provide context regarding cancer stage (early vs. late disease), leading to a lack of information on whether these mutations occur exclusively in early- or late-stage tumors, and also how external factors such as age and race may influence these mutations. However, further validation using cBioPortal, which provides mutation data linked to cancer stages and clinical outcomes, will be crucial to determine whether these frequently mutated genes are enriched in early-stage or late-stage prostate cancer.

The same gene symbols were evaluated using TCGA data via cBioPortal, and those genes with an alteration frequency of 5% or higher were selected, resulting in the OncoPrint shown in **Multimedia Appendix 1**. Visual inspection of selected genes including DLGAP2, PCDH17, CALB1, and VGLL2 revealed distinct patterns of genetic alterations across prostate cancer stages, ranging from T2A to T2C (early) to T3A to T4 (late). Notably, DLGAP2 and VGLL2 exhibited prominent deep deletions represented in blue, found across various stages but predominantly in early-stage prostate cancer, suggesting a potential tumor suppressor role. In contrast, CALB1 and PENK displayed consistent amplifications represented in red, especially in late-stage samples, indicating a possible association with disease progression or aggressive traits. According to the OncoPrint results, race and age at diagnosis did not show a strong correlation with gene alterations, as their distributions appeared random and inconsistent across cancer stages. These stage-associated alteration patterns reinforce the likelihood that these genes may contribute differentially to prostate cancer. This visual stratification, when aligned with earlier COSMIC mutation rates, supports prioritizing these genes for further analysis based on mutation type, frequency, and stage-specific behavior in clinical samples.

MLP Performance

Based on conventional methods of consideration and preliminary experiments, several MLP architectures have been selected to be tested on numerical data. The structures used consist of two main categories, namely MLPs with one hidden layer and two hidden layers. For the one-hidden-layer architecture, the number of neurons used varied between 8, 16, and 32, with ReLU or sigmoid activation functions, and dropouts between 0.2 and 0.4 to prevent overfitting. Meanwhile, the

architecture of the two hidden layers was designed with a combination of neuron counts such as 16-8 and 32-16, using consistent activation functions between layers (ReLU or sigmoid), as well as customized dropouts in each layer. The selection of these combinations aims to evaluate the effect of network depth, number of neurons, activation function, and dropout on the classification performance of the model.

To identify weaknesses in model performance, this study uses 3-fold cross validation method on training data with stratification on labels to ensure equal label distribution between Train and Test data. Prior to stratification, all features were scaled using scikit-learn StandardScaler() function to equalize the scale of values between features and speed up convergence during training.

Experiments in this study totaled 150 scenarios trained in a maximum of 100 epochs, the number of scenarios came from a combination of five feature selection datasets, three data splitting schemes, and 10 model architectures. All scenarios are trained using an early stopping mechanism with a patience value of 10 taken from 10% of the number of epochs, training will be stopped automatically if there is no performance improvement on the validation data for a specified number of epochs. In addition, all optimizers used have a learning rate of 0.001 consistently across all training scenarios, to maintain the stability of the reweighting process.

From the various scenarios, presented in Tables below, Table 3 presents the Top 5 best performing models out of the total 150 experiment scenarios and Table 4 presents 5 best performing models out of each individual feature counts ranging from 60 to 100.

Table 3. Top 5 performing models out of 150 experiment scenarios.

Features	Data Split	CV Acc	CV AUC	CV F1	Test ACC	Test Loss	Test AUC	Test F1	Asian Acc	Asian F1	Black Acc	Black F1
90	80/20	95%	99%	95%	98%	15%	100%	98%	25%	29%	43%	49%
60	80/20	95%	98%	95%	93%	20%	98%	92%	33%	33%	42%	47%
100	90/10	95%	98%	95%	95%	16%	99%	95%	8%	13%	47%	52%
70	80/20	95%	99%	95%	95%	15%	100%	95%	8%	13%	48%	54%
100	80/20	95%	99%	94%	99%	14%	100%	99%	8%	13%	47%	52%

Table 4. Top 5 performing model utilizing each feature count.

Features	Data Split	CV Acc	CV AUC	CV F1	Test ACC	Test Loss	Test AUC	Test F1	Asian Acc	Asian F1	Black Acc	Black F1
60	80/20	95%	98%	95%	93%	20%	98%	92%	33%	33%	42%	47%
60	90/10	93%	98%	93%	95%	12%	100%	95%	17%	17%	43%	49%
60	80/20	92%	96%	92%	94%	21%	99%	94%	17%	24%	43%	50%
60	70/30	91%	95%	91%	93%	24%	98%	93%	25%	29%	32%	35%
60	70/30	89%	95%	88%	94%	18%	99%	94%	25%	29%	40%	45%

70	80/20	95%	99%	95%	95%	15%	100%	95%	8%	13%	48%	54%
70	90/10	94%	98%	94%	98%	18%	100%	98%	8%	3%	38%	42%
70	80/20	91%	94%	91%	86%	49%	83%	84%	25%	29%	52%	58%
70	90/10	90%	96%	90%	95%	22%	100%	95%	0%	0%	33%	36%
70	90/10	90%	95%	89%	100%	28%	100%	100%	8%	13%	42%	48%
80	80/20	95%	99%	94%	92%	24%	98%	91%	25%	33%	42%	48%
80	90/10	93%	98%	93%	95%	12%	99%	95%	8%	3%	33%	37%
80	90/10	91%	96%	91%	98%	19%	100%	98%	0%	0%	38%	42%
80	90/10	91%	97%	90%	86%	33%	95%	83%	33%	42%	58%	63%
80	90/10	91%	96%	90%	76%	38%	99%	66%	83%	76%	83%	76%
90	80/20	95%	99%	95%	98%	15%	100%	98%	25%	29%	43%	49%
90	90/10	94%	98%	94%	100%	11%	100%	100%	8%	3%	35%	39%
90	90/10	92%	97%	92%	98%	16%	100%	98%	0%	0%	35%	39%
90	70/30	90%	94%	90%	98%	16%	100%	98%	8%	13%	38%	45%
90	80/20	89%	96%	89%	93%	29%	97%	92%	25%	33%	43%	50%
100	90/10	95%	98%	95%	95%	16%	99%	95%	8%	13%	47%	52%
100	80/20	95%	99%	94%	99%	14%	100%	99%	8%	13%	47%	52%
100	90/10	93%	96%	93%	98%	13%	99%	98%	0%	0%	37%	41%
100	90/10	92%	98%	92%	100%	12%	100%	100%	8%	3%	37%	40%
100	90/10	92%	98%	91%	76%	37%	95%	66%	83%	76%	83%	76%

Discussion

Principal Results

This paper proposes a race-based prostate cancer stage detection framework, which distinguishes between early and late stages, by utilizing bioinformatics feature selection (DMP) and statistical (RFE) approaches on patients' DNA methylation data. This framework is designed to minimize the number of features (CpG sites) used, to build a robust and accurate MLP model in classifying prostate cancer stages. By minimizing the number of features required, this approach provides several advantages for both patients and hospitals. These include reduced implementation costs due to the absence of large computational resources, as well as allowing patients to target specific CpG sites in the NGS process, so that the NGS process on irrelevant CpG sites can be eliminated. The generalization of the model is race-based due to the nature of DNA methylation data that is highly affected by disparities in lifestyle, genetic inheritance, and habits that vary by race [38–40].

In the labeling process, the distribution of prostate cancer stages in the sample data shows that most patients are diagnosed at stage 2, with 362 out of a total of 415 samples. Stage 1 accounted for only 21 samples, while stages 3 and 4 accounted for 39 and 82 samples respectively. This imbalance in distribution, particularly the high number of stages 2 and 4 compared to stage 1, may indicate a delay in early detection of prostate cancer in white individuals, where most prostate cancer cases are found at intermediate to advanced stages. Therefore, there is a need for accurate

prostate cancer staging detection methods to reduce the mortality rate of prostate cancer patients due to delayed or misdiagnosis.

To test the performance of the MLP-based classification framework, a series of experiments were conducted using data with varying feature count configurations, but with variations in the proportion of training and testing data (Data Split) and racial distribution. The results showed that the performance of the MLP model was generally quite high, as indicated by the accuracy, F1-score, and AUC values on the test data that approached or reached the maximum number (1.00) in some configurations. However, a more in-depth analysis of the model performance by racial group revealed significant inequalities.

Further analysis revealed that, although this experiment focused on configuring the number of features, the same trend was also observed in testing the model on a smaller number of features in a separate dataset. The cross-validation accuracy (CV Acc) values tend to stabilize in the range of 89% - 95% with a reduction in the number of features, indicating that the more information available, the more influential it is for the model to recognize relevant predictive patterns. However, some configurations with a limited number of features still show high accuracy. This phenomenon suggests that not only the quantity of features is important, but also the quality and biological relatedness of the selected features. In the context of DNA methylation data, there may be deep biological connections between CpG sites that play a role in specific biological processes, such as regulation of gene expression in certain tissues or races. The selected features may represent functionally interconnected biological networks that, despite their limited number, are still informative enough to maintain model performance.

Based on the model performance results, the disparity in performance between racial groups is apparent when evaluation metrics such as accuracy and F1-score are compared on a per-race basis. While the model can achieve high accuracy and F1-score overall, the performance for minority races such as Asian and black is much lower. For example, in one of the configurations where the overall test accuracy reached 99%, the accuracy for Asian race was only recorded at 8% with an F1-score of 0.13, while black race only achieved an accuracy of about 47% and an F1-score of 0.52. This disparity indicates that MLP models tend to only recognize patterns originating from certain racial groups, most likely the majority group, and fail to generalize the prediction results to other groups. One of the main causes of this bias is thought to come from the imbalance in the number of samples between races. In the data used, the white race has a relatively high number of samples (63 samples), followed by the black race (60 samples), while the Asian race consists of only 12 samples. This imbalance causes the model to receive more information during the training process from the majority group (white race), thus being able to recognize their patterns better. In addition, the feature selection process conducted earlier was most likely based on the majority group, causing the selected features to be less relevant for minority races. This leads to the possibility that the biological characteristics captured in the DNA methylation profiles of each race may produce

different patterns. In the context of this study, the CpG site features selected as the basis for prostate cancer stage classification appear to be more representative of the white group used for feature selection data and MLP model training, causing information on specific biomarkers in other races to be underrepresented. These biological differences may reflect interconnected genetic, epigenetic, environmental and social factors that influence the methylation patterns of each race. The impact of this imbalance is not only limited to the performance of MLP models but also reaches wider areas of medical practice and public health. If predictive models are built based on data that are not racially inclusive, the risk of bias in diagnosis, prognosis and clinical decision-making is higher, especially for racial minority. Therefore, the results of this study emphasize the importance of population representation aspects in DNA methylation data, as well as the need for methods that consider biological and demographic diversity in every stage of building artificial intelligence models for biomedical applications.

Limitations

This study has several limitations that should be addressed in future research. Most notably, the model was trained on a racially homogeneous dataset, which limits its generalizability to other populations. Our findings suggest that DNA methylation patterns varies significantly across racial groups, indicating that feature selection and model training must be performed on race-specific datasets to achieve accurate and fair classification outcomes. Without this, diagnostic models risk underperforming or misclassifying in underrepresented populations. Additionally, publicly available data was used which introduces challenges such as missing values and potential outdated annotations, which may influence model robustness. Finally, the study is limited to single-omics data; integrating multi-omics approaches could offer a more comprehensive view of prostate cancer progression across racial subgroups. Future work should prioritize race-aware feature selection approaches, stratified training pipelines, and advanced techniques such as domain adaptation to enhance both performance and equity in biomarker discovery.

Conclusions

This study presents a race-aware MLP model for prostate cancer stage classification using DNA methylation data, with enhanced feature selection through DMP analysis and RFE. Validations to external sources (COSMIC and cBioPortal) reinforced the credibility of our feature selection process, revealing important gene alteration patterns within the dataset. Results show that models built and tested within the White cohort achieved high performance (average testing accuracy >95%), but perform poorly when applied to Asian and African American samples (testing accuracy <50%). These findings reveal clear racial disparities in DNA methylation profiles and emphasize the critical need for race-specific feature selection and modeling strategies while demonstrating the effectiveness of using a minimal feature set to capture significant biomarkers while avoiding noise introduced by population heterogeneity.

Acknowledgements

This research is funded by Universitas Multimedia Nusantara Research Department under contract number ---.

Data Availability

This study utilized publicly available data from the UCSC Xena Browser. However, recent updates to the repository have resulted in missing components within the phenotype dataset. To address this, the complete phenotype dataset used in this research is provided in **Multimedia Appendix 2**.

Ethical Considerations

-

Author's Contributions

David Agustriawan, Moeljono Widjaja, Marlinda Vasty Overbeek, Muhammad Imran Ahmad, Srinivasulu Yerukala Sathipati participated in the concept, design, and led the critical revision of the manuscript for important intellectual content. Adithama Mulia participated in the drafting of the manuscript. Adithama Mulia, Vincent Kurniawan, Jheno Syechlo contributed to statistical analysis. David Agustriawan contributed to funding acquisition. David Agustriawan contributed to administrative, technical, or material support. David Agustriawan, Moeljono Widjaja, Marlinda Vasty Overbeek participated in supervision. All authors contributed to acquisition, analysis, or interpretation of data.

Conflicts of Interest

None Declared.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021 May 4;71(3):209–249. doi: 10.3322/caac.21660
2. Cook MB, Beachler DC, Parlett LE, Cochetti PT, Finkle WD, Lanes S, Hoover RN. Testosterone Therapy in Relation to Prostate Cancer in a U.S. Commercial Insurance Claims Database. *Cancer Epidemiology, Biomarkers & Prevention* 2020 Jan 1;29(1):236–245. doi: 10.1158/1055-9965.EPI-19-0619
3. Wang M, Chi G, Bodovski Y, Holder SL, Lengerich EJ, Wasserman E, McDonald AC. Temporal and spatial trends and determinants of aggressive prostate cancer among Black and White men with prostate cancer. *Cancer Causes & Control* 2020 Jan 15;31(1):63–71. doi: 10.1007/s10552-019-01249-0
4. Albertsen PC. Competing Risk Analysis of Men Aged 55 to 74 Years at Diagnosis Managed Conservatively for Clinically Localized Prostate Cancer. *JAMA* 1998 Sep 16;280(11):975. doi: 10.1001/jama.280.11.975
5. Atiq M, Chandran E, Karzai F, Madan RA, Aragon-Ching JB. Emerging treatment options for prostate cancer. *Expert Rev Anticancer Ther* 2023 Jun 3;23(6):625–631. doi: 10.1080/14737140.2023.2208352

6. Gao J, Shi W, Wang J, Guan C, Dong Q, Sheng J, Zou X, Xu Z, Ge Y, Yang C, Li J, Bao H, Zhong X, Cui Y. Research progress and applications of epigenetic biomarkers in cancer. *Front Pharmacol* 2024 Apr 12;15. doi: 10.3389/fphar.2024.1308309
7. Sherif ZA, Ogunwobi OO, Ressim HW. Mechanisms and technologies in cancer epigenetics. *Front Oncol* 2025 Jan 7;14. doi: 10.3389/fonc.2024.1513654
8. Ou W, Zhang X-X, Li B, Tuo Y, Lin R-X, Liu P-F, Guo J-P, Un H-C, Li M-H, Lei J-H, Gao X-J, Zheng F-F, Chen L-W, Long L-L, Wang Z-R. Integrated proteogenomic characterization of localized prostate cancer identifies biological insights and subtype-specific therapeutic strategies. *Nat Commun* 2025 Apr 3;16(1):3189. doi: 10.1038/s41467-025-58569-w
9. Stopsack KH, Su XA, Vasselkiv JB, Graff RE, Ebot EM, Pettersson A, Lis RT, Fiorentino M, Loda M, Penney KL, Lotan TL, Mucci LA. Transcriptomes of Prostate Cancer with *TMPRSS2:ERG* and Other ETS Fusions. *Molecular Cancer Research* 2023 Jan 3;21(1):14–23. doi: 10.1158/1541-7786.MCR-22-0446
10. Kwon W-A, Joung JY. Precision Targeting in Metastatic Prostate Cancer: Molecular Insights to Therapeutic Frontiers. *Biomolecules* 2025 Apr 27;15(5):625. doi: 10.3390/biom15050625
11. Brady L, Kriner M, Coleman I, Morrissey C, Roudier M, True LD, Gulati R, Plymate SR, Zhou Z, Birditt B, Meredith R, Geiss G, Hoang M, Beechem J, Nelson PS. Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. *Nat Commun* 2021 Mar 3;12(1):1426. doi: 10.1038/s41467-021-21615-4
12. Li M, Xia Z, Wang R, Xi M, Hou M. Unveiling DNA methylation: early diagnosis, risk assessment, and therapy for endometrial cancer. *Front Oncol* 2025 Jan 20;14. doi: 10.3389/fonc.2024.1455255
13. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol* 2021 Nov;82(11):801–811. doi: 10.1016/j.humimm.2021.02.012
14. Zhong Y, Xu F, Wu J, Schubert J, Li MM. Application of Next Generation Sequencing in Laboratory Medicine. *Ann Lab Med* 2021 Jan 1;41(1):25–43. doi: 10.3343/alm.2021.41.1.25
15. Wang Z, Yuan X, Nie Y, Wang J, Jiang G, Chen K. Next-Generation Sequencing vs. Clinical-Pathological Assessment in Diagnosis of Multiple Lung Cancers: A Systematic Review and Meta-Analysis. *Thorac Cancer* 2025 Mar 21;16(6). doi: 10.1111/1759-7714.70039
16. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics* 2022 Jul 25;16(1):26. doi: 10.1186/s40246-022-00396-x
17. Li S, Tollefsbol TO. DNA methylation methods: Global DNA methylation and methylomic analyses. *Methods* 2021 Mar;187:28–43. doi: 10.1016/j.ymeth.2020.10.002
18. Ehrlich M. Dna Hypomethylation In Cancer Cells. *Epigenomics* 2009 Dec 3;1(2):239–259. doi: 10.2217/epi.09.33
19. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, Aggarwal R, Playdle D, Liao A, Alumkal JJ, Das R, Chou J, Hua JT, Barnard TJ, Bailey AM, Chow ED,

- Perry MD, Dang HX, Yang R, Moussavi-Baygi R, Zhang L, Alshalalfa M, Laura Chang S, Houlahan KE, Shiah Y-J, Beer TM, Thomas G, Chi KN, Gleave M, Zoubeydi A, Reiter RE, Rettig MB, Witte O, Yvonne Kim M, Fong L, Spratt DE, Morgan TM, Bose R, Huang FW, Li H, Chesner L, Shenoy T, Goodarzi H, Asangani IA, Sandhu S, Lang JM, Mahajan NP, Lara PN, Evans CP, Febbo P, Batzoglou S, Knudsen KE, He HH, Huang J, Zwart W, Costello JF, Luo J, Tomlins SA, Wyatt AW, Dehm SM, Ashworth A, Gilbert LA, Boutros PC, Farh K, Chinnaiyan AM, Maher CA, Small EJ, Quigley DA, Feng FY. The DNA methylation landscape of advanced prostate cancer. *Nat Genet* 2020 Aug 13;52(8):778–789. doi: 10.1038/s41588-020-0648-8
20. Ramchoun H, Amine M, Idrissi J, Ghanou Y, Ettaouil M. Multilayer Perceptron: Architecture Optimization and Training. *International Journal of Interactive Multimedia and Artificial Intelligence* 2016;4(1):26. doi: 10.9781/ijimai.2016.415
 21. Gupta S, Kumar M. Prostate Cancer Prognosis Using Multi-Layer Perceptron and Class Balancing Techniques. 2021 Thirteenth International Conference on Contemporary Computing (IC3-2021) New York, NY, USA: ACM; 2021. p. 1–6. doi: 10.1145/3474124.3474125
 22. Lorencin I, Anđelić N, Španjol J, Car Z. Using multi-layer perceptron with Laplacian edge detector for bladder cancer diagnosis. *Artif Intell Med* 2020 Jan;102:101746. doi: 10.1016/j.artmed.2019.101746
 23. Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth* 2021;4:1–11. doi: 10.1016/j.ceh.2020.11.002
 24. Lee S-J, Tseng C-H, Lin GT –R., Yang Y, Yang P, Muhammad K, Pandey HM. A dimension-reduction based multilayer perception method for supporting the medical decision making. *Pattern Recognit Lett* 2020 Mar;131:15–22. doi: 10.1016/j.patrec.2019.11.026
 25. Liu M, Li L, Wang H, Guo X, Liu Y, Li Y, Song K, Shao Y, Wu F, Zhang J, Sun N, Zhang T, Luan L. A multilayer perceptron-based model applied to histopathology image classification of lung adenocarcinoma subtypes. *Front Oncol* 2023 May 18;13. doi: 10.3389/fonc.2023.1172234
 26. Abdollahi H, Mofid B, Shiri I, Razzaghdoust A, Saadipoor A, Mahdavi A, Galandooz HM, Mahdavi SR. Machine learning-based radiomic models to predict intensity-modulated radiation therapy response, Gleason score and stage in prostate cancer. *Radiol Med* 2019 Jun 3;124(6):555–567. doi: 10.1007/s11547-018-0966-4
 27. Hartenstein A, Lübke F, Baur ADJ, Rudolph MM, Furth C, Brenner W, Amthauer H, Hamm B, Makowski M, Penzkofer T. Prostate Cancer Nodal Staging: Using Deep Learning to Predict 68Ga-PSMA-Positivity from CT Imaging Alone. *Sci Rep* 2020 Feb 25;10(1):3398. doi: 10.1038/s41598-020-60311-z
 28. Eissa NS, Khairuddin U, Yusof R. A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on DNA methylation. *BMC Bioinformatics* 2022 Dec 11;23(1):273. doi: 10.1186/s12859-022-04815-7

29. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, Zhu J, Haussler D. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol. Nature Research*; 2020. p. 675–678. PMID:32444850
30. Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. *Bioinformatics* 2019 Oct 1;35(19):3786–3793. doi: 10.1093/bioinformatics/btz134
31. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012 Jul 29;13(7):484–492. doi: 10.1038/nrg3230
32. Stephen B. Edge, David R. Byrd, Carolyn C. Compton, April G. Fritz, Frederick L. Greene. *AJCC Cancer Staging Manual*. 7th ed. New York: Springer; 2010.
33. Borley N, Feneley MR. Prostate cancer: diagnosis and staging. *Asian J Androl* 2009 Jan 1;11(1):74–80. doi: 10.1038/aja.2008.19
34. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan J-B, Shen R. High density DNA methylation array with single CpG site resolution. *Genomics* 2011 Oct;98(4):288–295. doi: 10.1016/j.ygeno.2011.07.007
35. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 2019 Jan 8;47(D1):D941–D947. doi: 10.1093/nar/gky1015
36. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* 2012 May 1;2(5):401–404. doi: 10.1158/2159-8290.CD-12-0095
37. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signal* 2013 Apr 2;6(269). doi: 10.1126/scisignal.2004088
38. Guerrero-Preston R, Lawson F, Rodriguez-Torres S, Noordhuis MG, Pirini F, Manuel L, Valle BL, Hadar T, Rivera B, Folawiyo O, Baez A, Marchionni L, Koch WM, Westra WH, Kim YJ, Eshleman JR, Sidransky D. *JAK3* Variant, Immune Signatures, DNA Methylation, and Social Determinants Linked to Survival Racial Disparities in Head and Neck Cancer Patients. *Cancer Prevention Research* 2019 Apr 1;12(4):255–270. doi: 10.1158/1940-6207.CAPR-17-0356
39. Li Y, Pang X, Cui Z, Zhou Y, Mao F, Lin Y, Zhang X, Shen S, Zhu P, Zhao T, Sun Q, Zhang J. Genetic factors associated with cancer racial disparity – an integrative study across twenty-one cancer types. *Mol Oncol* 2020 Nov 24;14(11):2775–2786. doi: 10.1002/1878-0261.12799
40. Pepin ME, Ha C-M, Potter LA, Bakshi S, Barchue JP, Haj Asaad A, Pogwizd SM, Pamboukian S V., Hidalgo BA, Vickers SM, Wende AR. Racial and socioeconomic disparity associates with differences in cardiac DNA

methylation among men with end-stage heart failure. American Journal of Physiology-Heart and Circulatory Physiology 2021 May 1;320(5):H2066–H2079. doi: 10.1152/ajpheart.00036.2021

Abbreviations

MLP: Multi-Layer Perceptron
DMR: Differentially Methylated Region
DMP: Differentially Methylated Position
AR: Androgen Receptor
NGS: Next-Generation Sequencing
RFE: Recursive Feature Elimination
AJCC: American Joint Committee on Cancer
ROC: Receiver Operating Characteristic
AUC: Area Under the Curve
PSA: prostate-specific antigen
TCGA: The Cancer Genome Atlas
TCGA-PRAD: The Cancer Genome Atlas Prostate Adenocarcinoma
NGS: next-generation sequencing
MRI: magnetic resonance imaging
IMRT: intensity-modulated radiation therapy
T2W: T2-weighted
ADC: apparent diffusion coefficient
CNN: convolutional neural network
LNI: lymph node infiltration
UCSC: University of California, Santa-Cruz
CV: cross-validation