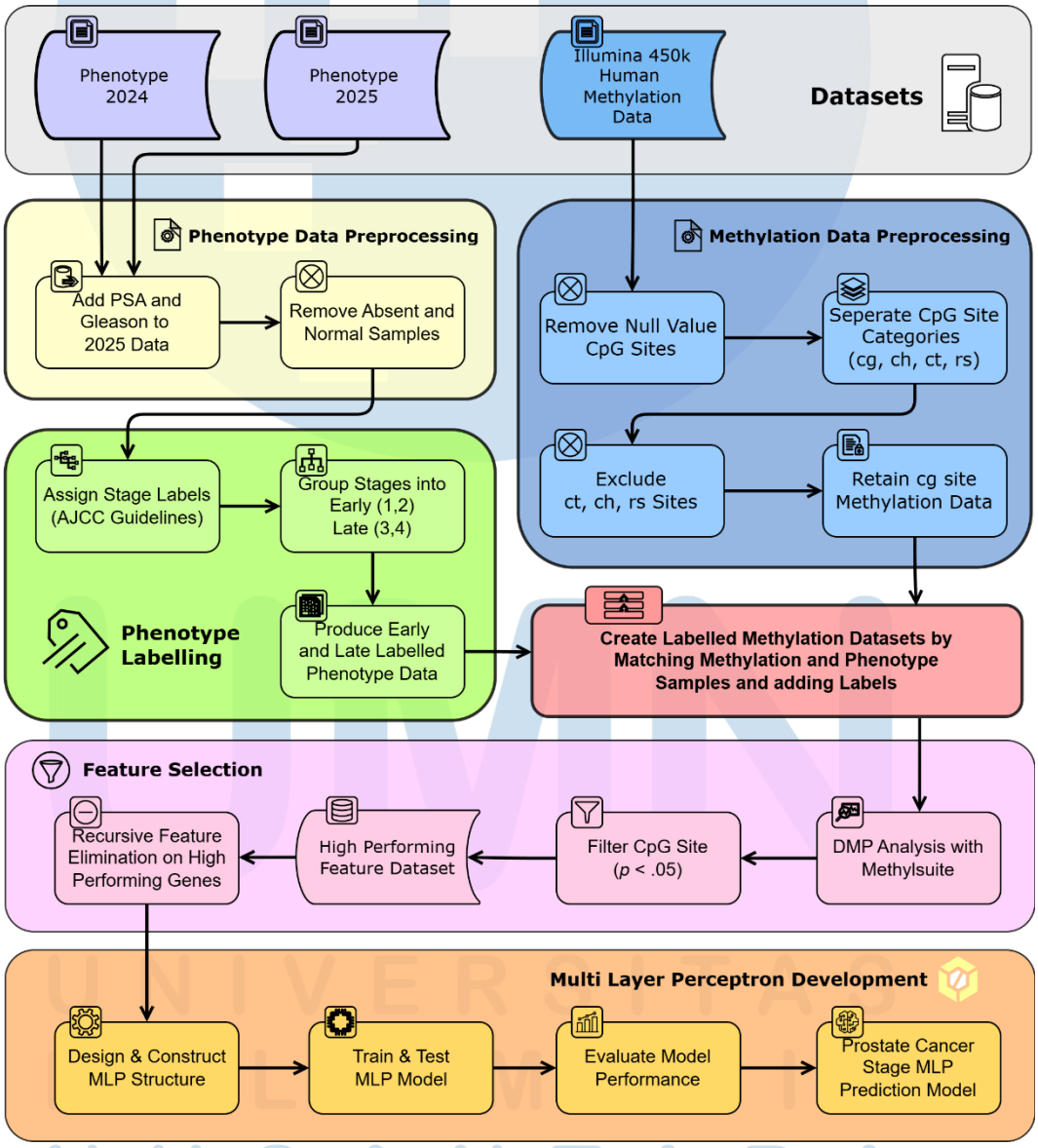
BAB III

METODE PENELITIAN

Penelitian ini menggunakan rancangan metode sesuai dengan Gambar 3. Perangkat yang digunakan selama menjalankan proses adalah Windows 11 OS, *Processor* Ryzen 7 5800H (8 Cores 16 Threads), GPU NVIDIA RTX 3070 Laptop 8GB, dan 32GB 3200MT/s RAM.



Gambar 3. Alur Penelitian.

3.1. Teknik Pengumpulan Data

Data dalam penelitian ini diambil dari UCSC Xenabrowser dataset dengan *cohort* GDC TCGA-PRAD. Data yang digunakan dalam penelitian ini merupakan data DNA Methylation Illumnina Human Methylation 450 dan phenotype yang di download pada 5 Februari 2025. Tambahan untuk data phenotype didapat dari repository yang sama pada awal tahun 2024.

3.2. Penanganan Missing Value

Dataset metilasi DNA bertipe illumina 450k terdiri dari sekitar 450 ribu CpG site berupa beta value berkisar dari 0 ~ 1 yang menandakan seberapa *hypermethylated* atau *hypomethylated* sebuah CpG site. Namun, dalam dataset terdapat sejumlah CpG site yang seluruh nilainya bernilai null di semua sampel dan secara parsial null pada sejumlah sampel, menandakan bahwa tidak tersedia data metilasi sama sekali untuk site tersebut. Dalam kasus ini, bukan hanya sel bernilai null yang diabaikan, melainkan seluruh kolom CpG site yang sepenuhnya kosong dikeluarkan dari proses seleksi fitur dan CpG site yang secara parsial kosong akan diganti dengan 0 untuk menandakan bahwa site tersebut tidak termetilasi, guna memastikan hanya fitur yang memiliki informasi valid yang dilibatkan dalam analisis selanjutnya.

3.3. Pemecahan Dataset

Dataset metilasi DNA tipe Illumina 450k terdiri dari empat jenis CpG site, yaitu cg, ct, ch, dan rs, yang masing-masing merepresentasikan; cg mentarget CpG site dinukleotida (sebuah sitosin yang diikuti oleh guanin), ch merupakan sitosin yang diikuti oleh basis selain guanin, ct meruapkan probe control yang berfungsi untuk memastikan kualitas dan normalisasi, bukan untuk mengukur metilasi pada CpG site, rs merupakan probe yang diketahui sebagai SNP (*Single Nucleotide Polymorphisms*) dan biasanya tidak digunakan untuk analisis sebab berpotensial untuk mengganggu hasil analisis dengan memberikan sinyal palsu [41]. Di antara keempat jenis tersebut, CpG site bertipe cg merupakan yang paling dominan, mencakup lebih dari 90% dari keseluruhan data. Terdapat nilai yang hilang sebagian di dalam sisa data cg yang tidak mencakup seluruh sampel, untuk

menangani hal ini, metode imputasi *mean* digunakan untuk mengkompensasi nilai yang hilang [42]. Sebaliknya, CpG site jenis ct, ch, dan rs memiliki tingkat missing value yang tinggi dengan tipe ch mengandung lebih dari 60% nilai null dan rs bahkan seluruhnya bernilai null. Karena tingginya proporsi nilai hilang pada ketiga tipe tersebut, dan juga untuk menjaga kualitas dan konsistensi data, penelitian ini hanya menggunakan fitur yang berasal dari CpG site bertipe cg, sedangkan fitur dari tipe ct, ch, dan rs dikeluarkan dari analisis (Tabel 1).

Tabel 1. Visualisasi pemecahan dataset berdasarkan jenis CpG site cg, ct, ch, rs.

	cg01	…	cgN	ctN	chN	rsN
Sample0	0.0231		0.896	null	null	null
Sample1	0.348		0.688	null	null	null
…	…		…	…	…	…
SampleN	Metilasi Sample N	…	Metilasi Sample N	Metilasi Sample N	Metilasi Sample N	Metilasi Sample N

Setelah didapatkan dataset cg, sampel dipisahkan berdasarkan setiap ras yang ada di data fenotipe. Ras dengan jumlah terbanyak digunakan sebagai data untuk *training* dan *testing*. Sampel ras yang tersisa digunakan sebagai validasi eksternal terhadap karakteristik data metilasi yang kemungkinan berbasis rasial.

3.4. Pelabelan Sampel

Penelitian ini dimulai dengan memberikan label kepada setiap sampel yang ada di dataset berdasarkan stadium kanker. Pelabelan dilakukan menggunakan manual pelabelan pada kanker prostat dari American Joint Committee on Cancer (AJCC) Cancer Staging Manual edisi ke-7 halaman 474 dilihat pada Gambar 4 [43].

UNIVERSITAS MULTIMEDIA NUSANTARA

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

Gambar 4. Instruksi pelabelan kanker prostat berdasarkan AJCC edisi ke-7 [43].

Terdapat 5 stadium stadium kanker prostat, pada proses pelabelan ini dipersingkat dengan menggabungkan stadium IIA dan IIB menjadi satu kelas yaitu stadium II mengurangkan total klasifikasi yang ada jadi 4 dan pelabelan dilakukan berdasarkan 4 stadium tersebut yang kemudian dipersingkat lebih lanjut dengan demikian, stadium 1 dan 2 diklasifikasikan sebagai stadium awal (*early*) sebab oleh sifat dari stadium tersebut yang dimana tumor belum menyebar ke area selain prostat sebagaimana stadium 3 dan 4 diklasifikasikan sebagai stadium akhir (*late*) sebab pada stadium tersebut tumor telah tersebar ke kelenjar getah bening (*lymph node*) atau bermetastasis [44], [45], [46]. Perubahan skema label ini juga dipengaruhi oleh ketergantungan metod analisis DMPs yang terbatas untuk fitur biner. Dengan merujuk beberapa kriteria yang membedakan stadium kanker di AJCC Cancer Staging Manual; T (keberadaan Tumor), N (persebaran terhadap nodus limfa), M (metastasis), PSA dan Gleason, keterangan lebih lanjut mengenai

kriteria-kriteria tersebut terdapat dalam Tabel 2. Namun karena kriteria PSA (Prostate Specific Antigen) dan Gleason tidak terdapat di dataset fenotipe Xenabrowser di tahun 2025, penelitian ini memenuhi kriteria yang tidak lengkap itu dari dataset fenotipe Xenabrowser pada tahun 2024.

Tabel 2. Penjelasan label group AJCC [43].

Label AJCC	Keterangan
TX	Tumor primer tidak dapat dinilai
T0	Tidak ada bukti tumor primer
T1	Tumor yang secara klinis tidak terlihat, tidak dapat diraba dan tidak dapat dilihat dengan pencitraan
T1a	Temuan histologis insidental tumor pada 5% atau kurang dari jaringan yang direseksi
T1b	Temuan histologis insidental tumor pada lebih dari 5% jaringan yang direseksi
T1c	Tumor diidentifikasi dengan biopsi jarum
T2	Tumor yang ditemukan di dalam prostat
T2a	Tumor melibatkan setengah dari satu lobus atau kurang
T2b	Tumor melibatkan lebih dari setengah bagian dari satu lobus tetapi tidak pada kedua lobus
T2c	Tumor melibatkan kedua lobus
T3	Tumor meluas melalui kapsul prostat
T3a	Ekstensi ekstrakapsular (unilateral atau bilateral)
T3b	Tumor menyerang vesikula seminalis
T4	Tumor tumbuh atau menyerang struktur yang berdekatan selain vesikula seminalis.
NX	Noda limfus tidak dapat dinilai
N0	Tidak ada metastasis pada noda limfus
N1	Metastasis pada noda limfus
M0	Tidak ada metastasis jauh
M1	Terdapat metastasis jauh
M1a	Terdapat metastasis pada noda limfus
M1b	Terdapat metastasis pada tulang
M1c	Terdapat metastasis pada tempat lain atau tanpa penyakit ulang

3.5. Seleksi Fitur

Seleksi fitur dimulai dengan analisis DMPs pada seluruh gene kepada label guna mencari nilai metilasi yang outlier terhadap label. Analisis DMPs dijalankan

dengan *python package Methylsuite* menggunakan fungsi methylize.diff_meth_pos.diff_meth_pos yang menerima ID sampel, beta value, dan label. Fungsi tersebut mengembalikan hasil analisis DMPs yang terdiri dari kolom; cg pos, P-val. Kemudian dari hasil analisis diberikan batas potong pada P-value bernilai dibawah 0.05. Jumlah cg yang didapat dari analisis tersebut dijadikan sebagai fitur untuk model MLP. Demi meminimalisir jumlah fitur lebih lanjut diimplementasikan algoritma RFE dengan menggunakan algoritma SVM sebagai *classifier*. Proses RFE menghasilkan beberapa subset data yang digunakan untuk mengevaluasi batas minimum jumlah fitur, di mana fitur yang terlalu sedikit menyebabkan model kesulitan dalam membedakan antar kelas.

Untuk memvalidasi hasil dari seleksi fitur, dilakukan pengecekan mutase gen pada semua CpG site yang terseleksi. Pengecekan mutase gen dimulai dengan mengambil Illumina ID dari CpG site (contoh: cg00000029) dan di mencocokan Illumina ID dengan symbol gen yang tersedia dari *Infinium HumanMethylation450 v1.2 BeadChip Product Files*. Setelah mendapatkan list symbol gen, pengecekan mutase gen dapat dilakukan dengan mengunduh dataset eksternal mengenai gen knaker yang diketahui terdapat mutasi di kasus kanker prostat oleh *Catalouge Of Somatic Mutations In Cancer* (COSMIC), mencocokan simbol gen pada dataset dan menghitung persentase mutase individual gen dari jumlah sampel yang telah diuji dengan jumlah sampel yang tercatat mutase [47]. Validasi lebih lanjut dapat dilakukan pada website cbioportal.org dengan men-*query* simbol gen pada dataset *Prostate Adenocarcinoma* (TCGA, PanCancer Atlas) untuk memperoleh informasi lebih detail mengenai mutase gen per-stadium kanker prostat pada dataset sampel yang digunakan pada penelitian ini (TCGA) [48], [49], [50].

3.6. Perancangan Model MLP

Konstruksi model MLP dimulai dengan menentukan skema dari MLP yang ingin digunakan. Umumnya, digunakan satu hingga dua lapisan tersembunyi karena data numerik/tabular cenderung memiliki struktur yang lebih sederhana dibanding data citra atau teks. Jumlah neuron pada setiap lapisan biasanya ditentukan secara heuristik, misalnya dengan mengambil rata-rata dari jumlah neuron input dan

output, atau menggunakan pola menurun seperti 32-16-8. Fungsi aktivasi seperti ReLU umum digunakan pada lapisan tersembunyi untuk meningkatkan kemampuan representasi data non-linear, sementara fungsi aktivasi sigmoid atau softmax digunakan pada lapisan *output* sesuai dengan jenis klasifikasi, dalam kasus ini merupakan klasifikasi biner maka menggunakan fungsi aktivasi sigmoid untuk semua lapisan *output*. Pemilihan arsitektur akhir umumnya ditentukan melalui eksperimen dan validasi untuk menghindari overfitting dan memastikan performa model yang optimal.

Setelah menentukan skema model MLP training model dimulai dengan mempisahkan dataset menjadi data *training* dan *testing*. Data *training* dan *testing* didapat dari memisahkan dataset awal. Metode *cross validation* diimplementasikan kepada proses *training* untuk mencegah potensial terjadinya *overfitting* atau bias kepada data *training* sehingga memperluas generalisir model. Pemisahan dataset ini dilakukan dengan beberapa skala pada Tabel 3.

Tabel 3. Skema pembagian data pada MLP.

Training	Testing
70%	30%
80%	20%
90%	10%

Training model dilakukan secara bertahap dalam satuan *epoch* hingga proses dihentikan secara otomatis oleh mekanisme *early stopping*. Mekanisme ini menghentikan pelatihan apabila tidak terjadi peningkatan performa model pada data validasi setelah sejumlah *epoch* tertentu, sehingga dapat mencegah overfitting. Selama proses pelatihan, model dievaluasi menggunakan optimizer Adam. Seluruh skenario pelatihan menggunakan fungsi loss yang konsisten, yakni *binary crossentropy*, sesuai dengan sifat masalah klasifikasi biner yang dihadapi.

3.7. Evaluasi Model

Evaluasi model dilakukan menggunakan fungsi *classification_report*, dengan fokus pada harmonisasi antara metrik *f1-score*, *accuracy*, dan nilai AUC. Kinerja model dievaluasi berdasarkan hasil pada data training dan testing untuk

memastikan keandalan tanpa adanya overfitting atau underfitting. Selain itu, nilai AUC (*Area Under the Curve*) juga digunakan sebagai metrik tambahan untuk menilai kemampuan model dalam membedakan antar kelas secara keseluruhan.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA