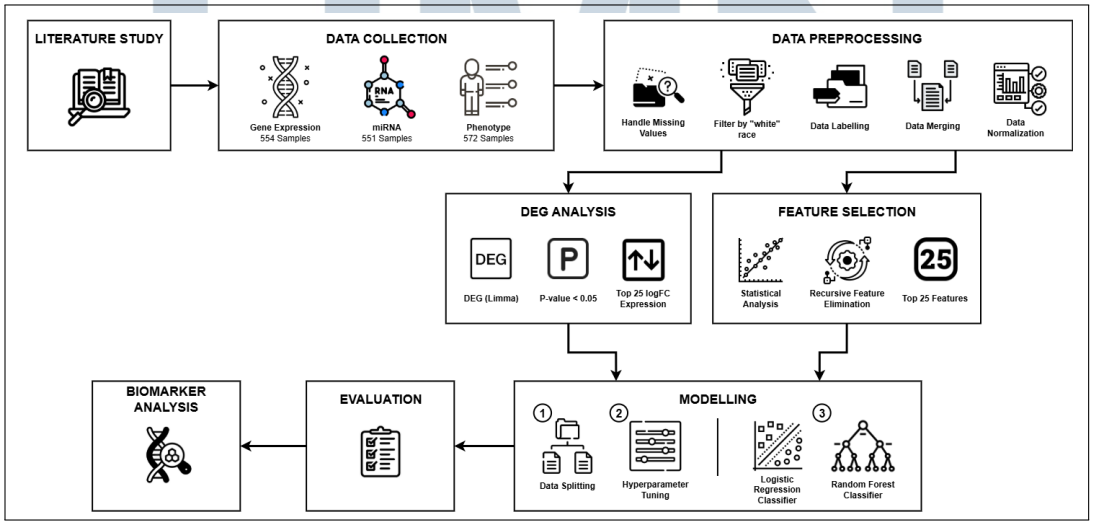
BAB 3
METODOLOGI PENELITIAN

3.1 Alur Kerja Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang dirancang secara sistematis dan terstruktur untuk mendukung pencapaian tujuan penelitian. Tahap awal yang dilakukan dalam penelitian ini adalah studi literatur, yang bertujuan untuk memperoleh pemahaman mendalam mengenai konteks permasalahan, teori-teori yang mendasari, serta pendekatan-pendekatan yang relevan dengan topik penelitian. Tahap selanjutnya adalah pengumpulan data, yang dilakukan melalui sumber terbuka yang terpercaya. Setelah data terkumpul, dilakukan proses pengolahan data untuk memastikan kualitas dan kelayakan data. Tahap selanjutnya adalah proses seleksi fitur untuk mendapatkan informasi yang paling relevan dan menyaring informasi yang redundan. Fitur-fitur yang telah terseleksi tersebut selanjutnya digunakan sebagai input dalam tahap pembangunan model klasifikasi. Setelah model didapatkan, dilakukan proses evaluasi terhadap model tersebut menggunakan berbagai metrik evaluasi. Tahap terakhir yang dilakukan adalah proses interpretasi dan analisis terhadap fitur-fitur yang terpilih oleh model dengan performa terbaik. Alur kerja penelitian dapat dilihat pada Gambar 3.1 untuk memberikan visualisasi yang lebih jelas.



Gambar 3.1. Diagram alur kerja penelitian

3.2 Spesifikasi Perangkat

Dalam pelaksanaan penelitian, terdapat beberapa perangkat yang mendukung kelancaran proses pengerjaan. Perangkat tersebut terbagi ke dalam dua kategori utama, yaitu perangkat keras dan perangkat lunak. Berikut adalah spesifikasi perangkat yang digunakan:

3.2.1 Perangkat Keras

1. Prosesor: 11th Gen Intel(R) Core(TM) i7-1165G @2.8Ghz.
2. Ram: 16 GB.
3. Memory: 512 GB.
4. Kartu Grafis: NVIDIA GeForce MX450.

3.2.2 Perangkat Lunak

1. Sistem Operasi: Windows 10 Pro 64-bit.
2. Bahasa Pemrograman: Python (versi 3.10.0), R(versi 4.2.1).
3. Editor Teks: Visual Studio Code, RStudio.
4. Web Browser: Google Chrome.

3.3 Studi Literatur

Studi literatur merupakan tahap awal yang penting dalam proses penelitian ini. Tujuan utama dari tahap ini adalah untuk memperoleh pemahaman mengenai permasalahan yang diteliti, teori-teori yang mendasarinya, serta pendekatan dan metode yang relevan dengan topik penelitian. Proses studi literatur dilakukan dengan menelusuri berbagai sumber informasi ilmiah seperti jurnal internasional, artikel konferensi, buku teks, serta penelitian-penelitian sebelumnya. Topik yang dikaji meliputi teori dasar mengenai *gene expression* dan *miRNA expression*, metode analisis seperti *limma*, teknik seleksi fitur, serta algoritma klasifikasi seperti *Logistic Regression* (LR) dan *Random Forest* (RF). Informasi yang diperoleh dari tahap ini menjadi dasar pengetahuan agar dapat merumuskan masalah yang tepat, menyusun

metodologi penelitian, merancang skenario eksperimen, serta memilih teknik dan parameter yang digunakan dalam penelitian.

3.4 Pengumpulan Data

Penelitian ini menggunakan dataset dari *Genomic Data Commons* (GDC) *The Cancer Genome Atlas Prostate Adenocarcinoma* (TCGA-PRAD) yang diakses melalui *Xena Browser* [37]. Dataset ini merupakan salah satu sumber terpercaya yang banyak digunakan dalam penelitian bioinformatika dan kanker, karena menyediakan data genomik dan klinis yang komprehensif untuk berbagai jenis kanker, termasuk kanker prostat. Dataset yang diambil terbagi menjadi tiga jenis yaitu, *gene expression RNAseq STAR - Counts*, *stem loop expression - miRNA Expression Quantification*, dan *phenotype*. Dataset yang digunakan dan detailnya disajikan pada Tabel 4.1

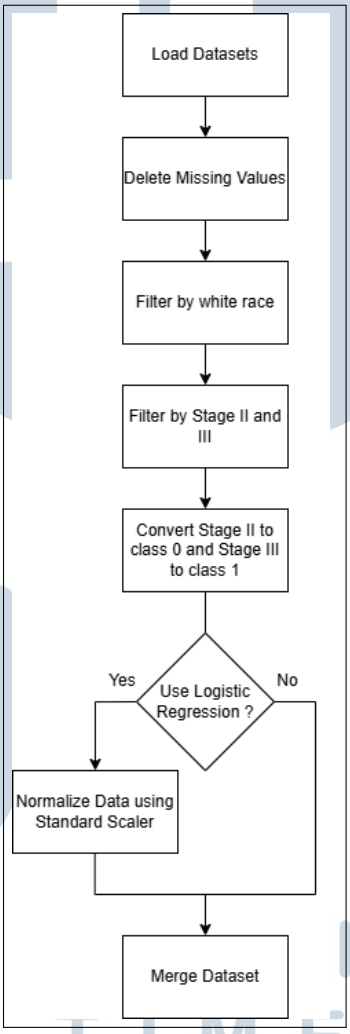
Tabel 3.1. Dataset Penelitian

Informasi	**RNAseq STAR - Counts**	**miRNA Expression**	**phenotype**
Jumlah Sampel	554	551	572
Jumlah Kolom	60.661 gen	1.882 miRNA	88 atribut
Versi Data	05-09-2024	05-09-2024	09-07-2024

Data RNA-seq yang digunakan dalam penelitian ini adalah STAR - Counts, yaitu jumlah bacaan mentah berdasarkan metode *Spliced Transcripts Alignment to a Reference* (STAR). Data RNA-seq tersebut beserta data miRNA telah melalui transformasi data dalam bentuk logaritmik log2(x + 1). Data *phenotype* merupakan data klinis yang terdiri dari 572 sampel pasien yang mencakup informasi penting seperti usia, status vital, ras, dan atribut lainnya. Data *phenotype* yang didapatkan tidak memiliki informasi mengenai stadium kanker sampel. Data *phenotype* yang diperoleh tidak memiliki informasi mengenai stadium kanker pada masing-masing sampel. Oleh karena itu, untuk menentukan stadium kanker, digunakan sistem penilaian TNM (Tumor, Node, Metastasis) yang informasinya tersedia dalam data tersebut. Penentuan label stadium klinis dilakukan oleh dokter spesialis urologi dari RSUD Dr. Saiful Anwar, Malang, yang terafiliasi dengan Program Studi Urologi, Fakultas Kedokteran, Universitas Brawijaya, Malang, menggunakan sistem TNM.

3.5 Praproses Data (Data Preprocessing)

Praproses data merupakan tahap penting dalam proses analisis data, khususnya dalam bidang bioinformatika dan pembelajaran mesin, karena memastikan bahwa data yang digunakan bersih, konsisten, dan relevan dengan tujuan penelitian. Langkah-langkah praproses data yang dilakukan dalam penelitian ini mengikuti alur seperti ditunjukkan pada Gambar 3.2 dan dijelaskan secara rinci sebagai berikut:



Gambar 3.2. Diagram alur kerja pra-proses data

Langkah awal yang dilakukan adalah memuat seluruh dataset yang digunakan. Proses pemuatan data dilakukan menggunakan bahasa pemrograman Python dan pustaka seperti `pandas` untuk mempermudah pengelolaan data. Langkah selanjutnya adalah melakukan proses pembersihan data untuk menghapus

nilai-nilai kosong yang dapat mengganggu proses analisis. Nilai kosong dapat berasal dari ketidakhadiran data pada proses sequencing atau pengumpulan data klinis. Dalam penelitian ini, fokus utama penghapusan data terletak pada kolom `Stage (Label)` atau informasi mengenai stadium kanker.

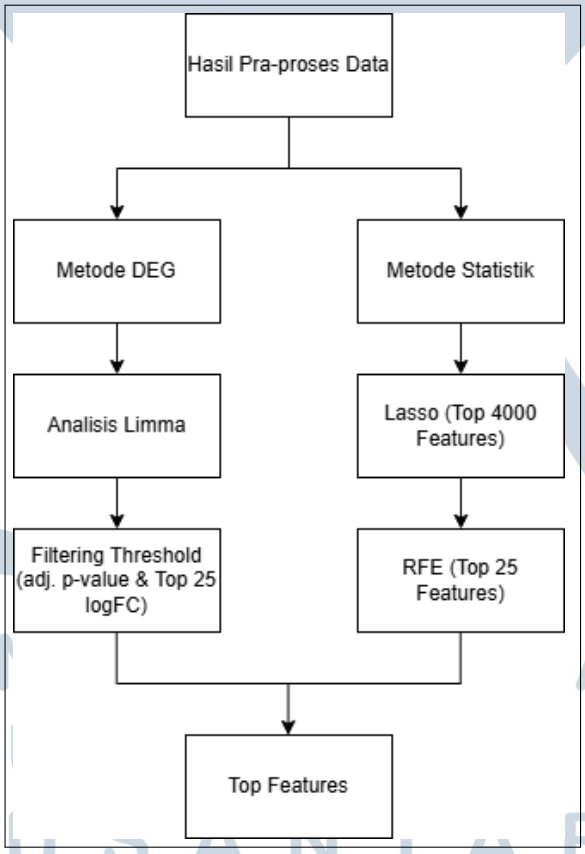
Setelah data sudah bersih dari nilai kosong, selanjutnya dilakukan *filtering* berdasarkan ras pasien. Dalam penelitian ini, hanya data pasien dengan ras kulit putih (*white*) yang digunakan karena ras ini mendominasi sekitar 83% populasi dataset. Hal ini dilakukan untuk mengurangi bias dan meningkatkan konsistensi hasil. Ekspresi gen dan progresi kanker dapat bervariasi antarkelompok etnis, sehingga penyaringan ini membantu dalam menstabilkan analisis. *filtering* data dilakukan pada kolom `race.demographic` yang berisi informasi mengenai ras pasien.

Pada data *phenotype* yang telah diberikan label oleh dokter spesialis urologi dari RSUD Dr. Saiful Anwar, informasi stadium kanker pasien bervariasi dari stadium I hingga IV. Pada penelitian ini, klasifikasi stadium kanker prostat difokuskan pada stadium II (*locally advanced cancer, early stage*), dan stadium III (*locally advanced cancer, late stage*). Sehingga pada tahap selanjutnya, dilakukan proses *filtering* terhadap kolom `Stage (Label)` untuk memperoleh data yang hanya mencakup sampel dengan stadium II dan III. Selanjutnya, data tersebut dikonversi menjadi format numerik menggunakan *label encoding*, di mana stadium II direpresentasikan sebagai 0 dan stadium III sebagai 1 ke dalam kolom yang bernama `Stage_Binary`.

Data yang sudah diproses tersebut selanjutnya akan digabungkan menjadi dua buah dataset yang berbeda. Dataset pertama merupakan hasil dari penggabungan antara data *gene expression* dan data *phenotype*. Dataset kedua merupakan hasil dari penggabungan antara data *miRNA expression* dan data *phenotype*. Penggabungan data tersebut dilakukan dengan metode inner join berdasarkan indeks sampel, di mana kolom yang diambil dari data *phenotype* hanyalah kolom `Stage_Binary`. Tahap terakhir dari pra-proses data adalah melakukan normalisasi data menggunakan *z-score standardization* guna menyamakan skala antar fitur. Proses normalisasi tersebut hanya dilakukan pada percobaan yang menggunakan algoritma *Logistic Regression*.

3.6 Seleksi Fitur

Seleksi fitur merupakan tahap penting dalam proses analisis data, khususnya pada data berdimensi tinggi seperti ekspresi gen dan miRNA. Alur kerja seleksi fitur dimulai setelah tahap praproses data selesai dilakukan. Tahap ini bertujuan untuk mengurangi jumlah fitur yang tidak relevan atau redundan, sehingga dapat meningkatkan kinerja model klasifikasi dan mengurangi kompleksitas komputasi. Dalam penelitian ini, seleksi fitur diterapkan dengan dua buah pendekatan yang berbeda. Pendekatan pertama menggunakan basis bioinformatika dengan analisis *Differentially Expressed Genes* (DEG). Pendekatan kedua menggunakan basis metode statistik. Kedua pendekatan ini dirancang untuk menghasilkan subset fitur yang paling relevan terhadap target klasifikasi dan akan dievaluasi lebih lanjut menggunakan model klasifikasi. Langkah-langkah seleksi fitur yang dilakukan dalam penelitian ini mengikuti alur seperti ditunjukkan pada Gambar 3.3.



Gambar 3.3. Diagram alur kerja seleksi fitur

3.6.1 Seleksi Fitur Berbasis DEG

Seleksi fitur berbasis *Differentially Expressed Genes* (DEG) merupakan pendekatan yang berfokus pada aspek biologis, di mana fitur yang terpilih diharapkan memiliki keterkaitan fungsional terhadap fenomena biologis yang sedang dikaji. Dalam penelitian ini, analisis DEG dilakukan menggunakan paket `limma` pada lingkungan R. Tahapan analisis meliputi transformasi dan normalisasi data, penerapan model linier untuk setiap fitur, serta uji signifikansi statistik menggunakan metode moderasi variansi.

Hasil dari penerapan metode limma akan menghasilkan berbagai nilai statistik yang akan digunakan untuk melakukan seleksi. Nilai-nilai tersebut termasuk nilai *log-fold change*, *p-value*, *adjusted p-value*, *average expression*, dan nilai statistik t. Dalam penelitian ini, setiap fitur dievaluasi berdasarkan nilai *adjusted p-value* serta nilai *log-fold change* yang menggambarkan besarnya perubahan ekspresi antar kelas. Fitur yang diambil harus memenuhi kriteria ambang batas *adjusted p-value* $< 0{,}05$. Kemudian dari fitur yang telah terseleksi, diambil 25 fitur dengan *absolute value* tertinggi nilai *log-fold change*. 25 fitur tersebut dipilih sebagai fitur signifikan yang akan digunakan pada tahap klasifikasi.

3.6.2 Seleksi Fitur Berbasis Metode Statistik

Pendekatan kedua yang diterapkan pada penelitian ini adalah pendekatan dengan metode statistik. Pendekatan ini difokuskan pada hubungan atau korelasi antara fitur (gen atau miRNA) dengan label kelas dari sudut pandang statistik dan komputasional, tanpa mempertimbangkan informasi biologis secara langsung. Pendekatan ini bertujuan untuk menyaring fitur-fitur yang memiliki kekuatan diskriminatif tinggi terhadap klasifikasi dua kelompok kelas secara statistik. Beberapa metode statistik yang digunakan dalam penelitian ini meliputi LASSO (*Least Absolute Shrinkage and Selection Operator*), dan *Recursive Feature Elimination* (RFE). Dalam pendekatan ini, metode seleksi fitur diterapkan dalam bentuk kombinasi dua metode tersebut. LASSO digunakan sebagai tahap awal pemfilteran fitur, kemudian dilanjutkan dengan metode RFE sebagai tahap penyempurnaan. 25 fitur terbaik akan diambil sebagai hasil dan digunakan untuk melakukan pembangunan model.

3.7 Pembangunan Model Klasifikasi

Pembangunan model klasifikasi dilakukan setelah proses seleksi fitur selesai dilaksanakan dan 25 fitur terbaik telah berhasil diperoleh dari masing-masing skenario seleksi fitur. Pada penelitian ini, dua algoritma pembelajaran mesin digunakan untuk membangun model klasifikasi, yaitu *Logistic Regression* dan *Random Forest* (RF). Kedua algoritma ini dipilih karena memiliki performa yang baik dalam menangani data berdimensi tinggi serta telah banyak digunakan dalam penelitian bioinformatika.

Sebelum model dibangun, data akan dibagi menjadi dua bagian, yaitu data pelatihan (*training*) dan data pengujian (*testing*) dengan rasio pembagian sebesar 80:20. Pembagian ini bertujuan untuk memisahkan data yang digunakan untuk membangun model dengan data yang digunakan untuk menguji generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya. Selanjutnya, dilakukan proses tuning hiperparameter menggunakan metode *Grid Search Cross Validation (Grid Search CV)* untuk menemukan kombinasi parameter terbaik pada masing-masing algoritma. *Grid Search CV* memungkinkan evaluasi performa model secara menyeluruh berdasarkan kombinasi parameter tertentu melalui validasi silang (*cross-validation*), dalam hal ini menggunakan *5-fold cross-validation*. Hal ini bertujuan untuk menghindari overfitting dan memastikan model yang dihasilkan memiliki performa yang stabil. Rincian ruang pencarian *hyperparameter* yang digunakan untuk model LR dan RF pada penelitian ini ditunjukkan pada Tabel 3.2 dan Tabel 3.3.

Tabel 3.2. Tuning Hyperparameter untuk LR

Hyperprameter	**Nilai yang dicoba**
C	0.001, 0.01, 0.1, 1, 10, 100
penalty	l2
solver	liblinear, lbfgs, sag, newton-cg

Tabel 3.3. Tuning Hyperparameter untuk RF

Hyperprameter	Nilai yang dicoba
n_estimators	100, 200, 500
max_depth	10, 20, 30
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	sqrt, log2

3.8 Evaluasi

Tahap evaluasi dilakukan untuk mengukur performa model klasifikasi yang dibangun dengan konfigurasi *hyperparameter* optimal. Penilaian performa model dilakukan dengan menggunakan data *testing* yang telah dipisahkan dari data *training* sebelumnya. Metrik utama yang digunakan dalam penelitian ini adalah *confusion matrix*, yang diikuti dengan beberapa metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *f1-score*. Metrik-metrik ini dipilih karena mampu memberikan gambaran menyeluruh mengenai kemampuan prediksi model, khususnya dalam konteks klasifikasi biner.

3.9 Validasi Eksperimen dengan Random Seed Berulang

Dalam membangun dan mengevaluasi model klasifikasi, penggunaan parameter acak seperti pemisahan data pelatihan dan pengujian (*train-test split*) serta inisialisasi model dapat menyebabkan variasi hasil yang cukup signifikan. Untuk memastikan bahwa performa model yang dihasilkan tidak tergantung pada satu nilai acak tertentu, penelitian ini menerapkan strategi *multiple random seed*. Pendekatan ini dilakukan dengan menjalankan tahap pembangunan dan evaluasi model secara berulang-ulang menggunakan nilai seed yang berbeda. Dalam penelitian ini, digunakan sebanyak lima nilai seed acak yang berbeda, yaitu seed 42 hingga 46. Untuk setiap seed, dilakukan proses pemisahan data, seleksi fitur, tuning hiperparameter, pelatihan model, serta evaluasi model menggunakan metrik-metrik evaluasi. Langkah ini bertujuan untuk mengurangi potensi bias akibat pemisahan data yang tidak representatif, serta memberikan gambaran yang lebih stabil dan reliabel terhadap performa model. Rata-rata dari seluruh metrik evaluasi pada setiap seed akan dihitung untuk memberikan nilai performa akhir dari setiap skenario eksperimen.

3.10 Skenario Eksperimen

Proses eksperimen pada penelitian ini akan dilakukan berdasarkan beberapa skenario yang ditentukan. Pembagian skenario dilakukan berdasarkan variasi dalam algoritma klasifikasi, metode seleksi fitur yang digunakan, serta jenis data yang digunakan. Kombinasi dari ketiga faktor tersebut menghasilkan total enam skenario eksperimen yang berbeda. Rincian masing-masing skenario disajikan secara lengkap pada Tabel 3.4.

Tabel 3.4. Daftar Skenario Seleksi Fitur dan Klasifikasi

Nama Skenario	Jenis Data	Metode Seleksi Fitur	Algoritma Klasifikasi
Skenario 1	Gene	DEG (limma)	Logistic Regression
Skenario 2	miRNA	DEG (limma)	Logistic Regression
Skenario 3	Gene	DEG (limma)	Random Forest
Skenario 4	miRNA	DEG (limma)	Random Forest
Skenario 5	Gene	LASSO + RFE	Logistic Regression
Skenario 6	miRNA	LASSO + RFE	Logistic Regression
Skenario 7	Gene	LASSO + RFE	Random Forest
Skenario 8	miRNA	LASSO + RFE	Random Forest

3.11 Analisis Biomarker

Analisis biomarker bertujuan untuk mengidentifikasi kandidat biomarker potensial yang memiliki kontribusi penting dalam membedakan stadium kanker prostat. Analisis dilakukan pada *gen* atau *miRNA* yang terpilih pada model dengan performa terbaik. Pada penelitian ini, proses analisis dilakukan melalui beberapa pendekatan. Pendekatan pertama dilakukan dengan cara melakukan studi literatur terhadap gen yang terpilih untuk mengidentifikasi keterkaitannya dengan kanker prostat. Pendekatan selanjutnya dilakukan dengan cara melakukan pengujian gen-gen secara individu menggunakan ROC terhadap stadium kanker prostat.