BAB 3
METODOLOGI PENELITIAN

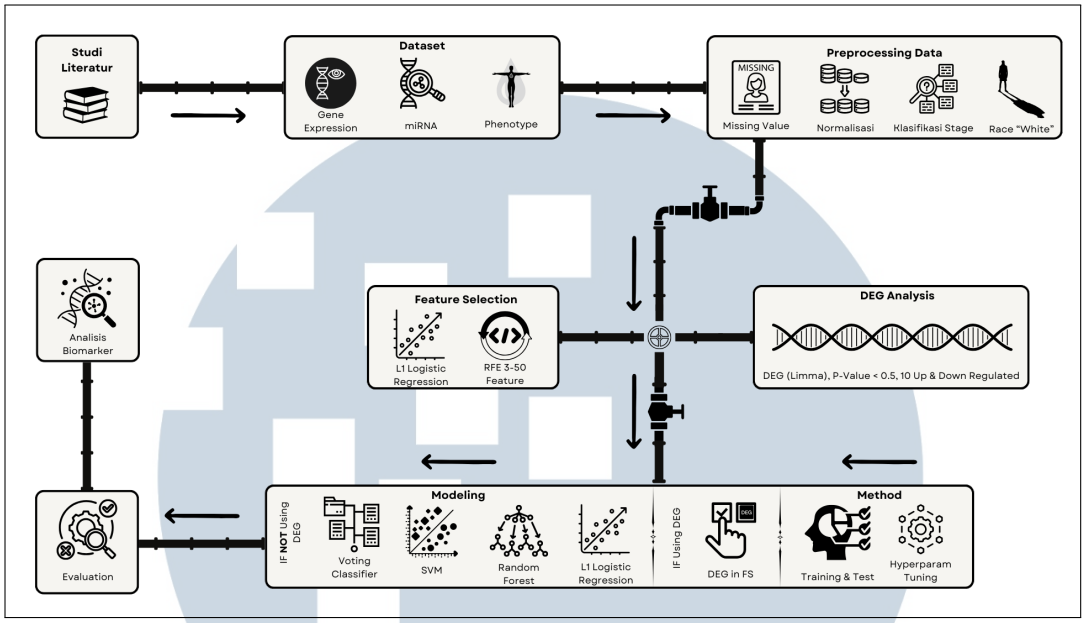
3.1 Gambaran Umum Penelitian

Penelitian ini berfokus pada upaya mendeteksi stadium kanker payudara dengan pendekatan komputasional berbasis data ekspresi genetik. Secara khusus, penelitian ini ditujukan untuk membedakan antara kanker payudara stadium 2 (*early stage*) dan stadium 3 (*advanced stage*), yang merupakan tahap krusial dalam menentukan arah penanganan medis dan prognosis pasien. Gambar 3.1 merupakan *research pipeline* untuk klasifikasi tahapan kanker payudara berdasarkan data ekspresi gen dan miRNA.

Data yang digunakan dalam penelitian ini bersumber dari *Genomic Data Commons* (GDC) *The Cancer Genome Atlas* (TCGA) untuk jenis kanker payudara (BRCA), yang diakses melalui platform UCSC Xena. Dataset ini mencakup data ekspresi gen dan miRNA dari ratusan pasien, serta informasi klinis berupa data *phenotype* yang memuat status stadium kanker.

Melalui pemrosesan data dan penerapan algoritma *machine learning*, penelitian ini bertujuan untuk mengembangkan model klasifikasi yang mampu mengidentifikasi stadium kanker secara akurat hanya berdasarkan data molekuler. Dengan pendekatan ini, dapat dihasilkan sistem pendukung keputusan yang lebih efisien dan minim invasif, sebagai pelengkap dari prosedur medis konvensional.

Langkah-langkah dalam penelitian meliputi pra-pemrosesan data, seleksi fitur dengan atau tanpa DEG dan algoritma *machine learning*, pelatihan model klasifikasi, evaluasi performa model, serta analisis biomarker. Seluruh tahapan ini dirancang untuk memastikan hasil klasifikasi yang tidak hanya akurat, tetapi juga stabil dan dapat diandalkan untuk diterapkan dalam konteks medis di masa depan.



Gambar 3.1. *Research Pipeline* untuk Klasifikasi Tahapan Kanker Payudara Berdasarkan Data Ekspresi Gen dan miRNA.

3.2 Spesifikasi Perangkat

Untuk melaksanakan penelitian ini, digunakan beberapa peralatan yang terbagi ke dalam dua kategori utama, yaitu perangkat keras dan perangkat lunak. Berikut adalah daftar perangkat keras dan perangkat lunak yang dipakai dalam penelitian ini.

1. Hardware

- Prosesor: AMD Ryzen 5 5600H with Radeon Graphics, 3.30 GHz.
- Memori: 16.0 GB RAM.
- Kartu Grafis (GPU): NVIDIA GeForce GTX 1650.
- Media Penyimpanan: 512 GB SSD.
- Server: Kaggle Accelerator GPU T4 x2.

2. Software

- Sistem Operasi: Windows 11 Home Single Language, versi 23H2.

- Web Browser: Google Chrome.
- Bahasa Pemrograman: LaTeX (Overleaf), Python, R.
- Text Editor: Visual Studio Code, RStudio.

3.3 Studi Literatur

Studi literatur merupakan langkah awal yang sangat penting dalam sebuah penelitian. Melalui studi literatur, peneliti mengumpulkan, mengevaluasi, dan memahami berbagai informasi dari sumber-sumber yang relevan dengan topik yang diteliti. Dengan melakukan hal ini, peneliti bisa menemukan celah atau kekurangan dari penelitian-penelitian sebelumnya serta memperkuat landasan teori dan kerangka konseptual yang digunakan. Selain itu, studi literatur juga sangat membantu dalam merumuskan masalah penelitian secara lebih jelas dan menyusun hipotesis yang tepat dan terarah.

3.4 Pengumpulan Data

Pengumpulan data pada penelitian ini menggunakan metode penelitian kuantitatif. Dataset yang digunakan dalam penelitian ini menggunakan dataset dari *Genomic Data Commons* (GDC) *The Cancer Genome Atlas* (TCGA) untuk jenis kanker payudara (BRCA), yang diakses melalui platform UCSC Xena [56]. Dataset yang diambil terbagi menjadi tiga jenis yaitu, *gene expression RNAseq STAR - FPKM-UQ* (60,661 *identifiers* X 1226 *samples*) versi 05-20-2024 , *stem loop expression - miRNA Expression Quantification* (1,882 *identifiers* X 1202 *samples*) versi 08-01-2024, dan *phenotype* (85 *identifiers* X 1255 *samples*) versi 09-07-2024. Untuk mengetahui gambaran mengenai dataset yang digunakan dalam penelitian ini, dapat dilihat pada Gambar 3.2 yang merupakan dataset *Gene Expression*, Gambar 3.3 yang merupakan dataset *miRNA Expression*, dan Gambar 3.4 yang merupakan dataset Phenotype.

Ensembl_ID	TCGA-D8-A146-01A	TCGA-AQ-A0Y5-01A	TCGA-C8-A274-01A	TCGA-BH-A0BD-01A	TCGA-B6-A1KC-01B	TCGA-AC-A62V-01A
ENSG00000000003.15	3.76770776	2.132741337	4.768009603	3.049735307	3.009311337	2.262162514
ENSG00000000005.6	1.759454145	0.146003451	0	1.071762669	0.18599337	0.308593869
ENSG00000000419.13	4.920541034	5.284151341	5.051694018	4.618796593	4.852972627	5.364064598
ENSG00000000457.14	2.612588407	2.428544424	3.356298779	2.680774425	2.381670993	1.128095396
ENSG00000000460.17	1.363395385	1.158337027	2.45817199	2.494236382	1.565840854	1.249809382
ENSG00000000938.13	1.920826827	1.088209455	0.981925699	1.638027485	0.78307943	1.308302621
ENSG00000000971.16	3.14745324	3.229680203	1.595073939	3.022172481	2.527145422	1.728007558
ENSG00000001036.14	3.72297962	4.749003215	3.433373455	3.817459538	3.926834421	5.573298778
ENSG00000001084.13	2.599508236	2.424545918	1.924061869	2.038646881	2.106180788	1.797386577
ENSG00000001167.14	3.939997565	3.951205437	4.234600951	3.883425316	4.635742785	2.720956963

Gambar 3.2. *Gene Expression RNAseq STAR - FPKM-UQ* Dataset

miRNA_ID	TCGA-D8-A146-01A	TCGA-AQ-A0Y5-01A	TCGA-C8-A1HI-01A	TCGA-A7-A0CD-01A	TCGA-5L-AAT1-01A	TCGA-BH-A0C0-01A
hsa-let-7a-1	12.81516715	12.96883789	13.26395076	12.87187945	13.86581539	12.5480326
hsa-let-7a-2	12.81605072	12.972606	13.24777909	12.87093887	13.86145328	12.52470102
hsa-let-7a-3	12.85501596	12.9724846	13.27008485	12.89658377	13.86704178	12.54656239
hsa-let-7b	14.7301584	14.49707982	13.84485204	14.93536386	14.60694951	14.47605727
hsa-let-7c	11.16962834	11.86926446	10.90436096	9.644680139	11.23780755	9.295809334
hsa-let-7d	8.8598326	7.23980758	9.067461254	7.84932823	8.316985445	9.544387533
hsa-let-7e	9.505781519	9.593766382	10.09700721	8.665823803	10.44448414	9.008172845
hsa-let-7f-1	10.68085335	11.53202895	12.52865535	9.653458484	12.4487946	10.666966
hsa-let-7f-2	10.72264842	11.47934769	12.53676441	9.730741411	12.47148351	10.59411557
hsa-let-7g	8.288113513	8.007486338	9.762910339	8.95974876	8.850126249	8.995297522

Gambar 3.3. *Stem Loop Expression - miRNA Expression Quantification* Dataset

sample	id	disease_type	case_id	submitter_id	primary_site
TCGA-BH-A0W3-01A	3c612e12-6de8-44fa-a095-805c45474821	Ductal and Lobular Neoplasms	3c612e12-6de8-44fa-a095-805c45474821	TCGA-BH-A0W3	Breast
TCGA-AR-A24V-01A	3cb06c7a-f2a8-448b-91a8-dd201bbf2ddd	Ductal and Lobular Neoplasms	3cb06c7a-f2a8-448b-91a8-dd201bbf2ddd	TCGA-AR-A24V	Breast
TCGA-E9-A1NE-01A	3d676bba-154b-4d22-ab59-d4d4da051b94	Ductal and Lobular Neoplasms	3d676bba-154b-4d22-ab59-d4d4da051b94	TCGA-E9-A1NE	Breast
TCGA-E9-A1NE-11A	3d676bba-154b-4d22-ab59-d4d4da051b94	Ductal and Lobular Neoplasms	3d676bba-154b-4d22-ab59-d4d4da051b94	TCGA-E9-A1NE	Breast
TCGA-AC-A8OQ-01A	dfaabd03-2d40-4422-b210-caf112ff4229	Ductal and Lobular Neoplasms	dfaabd03-2d40-4422-b210-caf112ff4229	TCGA-AC-A8OQ	Breast
TCGA-AC-A23C-01A	dfd0b7ba-c7d3-498e-b455-346301865452	Ductal and Lobular Neoplasms	dfd0b7ba-c7d3-498e-b455-346301865452	TCGA-AC-A23C	Breast
TCGA-C8-A12P-01A	abdc76db-f85e-4337-a57e-6d098789da03	Ductal and Lobular Neoplasms	abdc76db-f85e-4337-a57e-6d098789da03	TCGA-C8-A12P	Breast
TCGA-E9-A1N8-01A	ac075bc0-1b59-4557-beea-541694faee03	Ductal and Lobular Neoplasms	ac075bc0-1b59-4557-beea-541694faee03	TCGA-E9-A1N8	Breast
TCGA-E9-A1N8-11A	ac075bc0-1b59-4557-beea-541694faee03	Ductal and Lobular Neoplasms	ac075bc0-1b59-4557-beea-541694faee03	TCGA-E9-A1N8	Breast
TCGA-AN-A0XV-01A	ac18b3a3-8d52-4e35-8625-673171a7fd92	Ductal and Lobular Neoplasms	ac18b3a3-8d52-4e35-8625-673171a7fd92	TCGA-AN-A0XV	Breast

Gambar 3.4. *Phenotype* Dataset

3.5 Preprocessing Data

Preprocessing Data data merupakan proses yang kompleks dan mencakup berbagai tahapan penting. Proses ini dimulai dari penyiapan data, yang mencakup pembersihan data dari kesalahan atau duplikasi, pembersihan data kosong (*NaN*), klasifikasi, normalisasi untuk menyamakan skala data, standardisasi, *feature scaling*, hingga transformasi agar data lebih mudah dianalisis [57]. Tahap *preprocessing* data mencakup serangkaian proses penting yang bertujuan untuk menyiapkan data mentah agar siap digunakan dalam analisis lebih lanjut, berupa:

1. Penggabungan Dataset: Dataset dilakukan *transpose* pada data karena struktur awal data masih belum sesuai, yaitu posisi baris dan kolom

masih tertukar. Dataset ekspresi gen dikombinasikan dengan dataset *phenotype* untuk mengambil informasi penting seperti "*race.demographic*" dan "*ajcc_pathologic_stage.diagnoses*". Proses serupa juga diterapkan pada dataset *miRNA*, di mana informasi *phenotype* ditambahkan untuk melengkapi data biologis dengan data klinis yang relevan.

2. Penanganan *Missing Value*: Setelah proses penggabungan dataset, langkah selanjutnya adalah mengidentifikasi data yang memiliki nilai hilang (*NaN*). Baris yang mengandung nilai *NaN* akan dihapus untuk menjaga kualitas dan konsistensi data.

3. Klasifikasi *Stage*: Pada kolom *ajcc_pathologic_stage.diagnoses*, dilakukan proses *mapping* untuk mengelompokkan stadium kanker menjadi *stage* 1, 2, 3, 4, dan X. Setelah pemetaan, data dengan stadium 1, 4, dan X dihapus karena tidak termasuk dalam fokus penelitian. Penelitian ini secara khusus ditujukan untuk membedakan biomarker antara pasien stadium 2 dan stadium 3. Oleh karena itu, tahap selanjutnya adalah mengklasifikasikan *stage* 2 sebagai *early stage* dan *stage* 3 sebagai *advanced stage*.

4. Mengambil Ras *Race* White: Pada kolom *race.demographic* terdapat beragam kategori ras. Dalam penelitian ini, hanya data dengan ras putih (*White*) yang diambil. Hal ini didasarkan pada pertimbangan bahwa beberapa studi sebelumnya menunjukkan adanya hubungan yang signifikan antara ekspresi biomarker kanker payudara dengan kelompok ras putih non-Hispanik. Selain itu, pemilihan ras yang homogen juga bertujuan untuk mengurangi variabilitas dan meningkatkan keakuratan analisis.

5. Standardisasi dan *Feature Scaling* Dalam machine learning, sering kali data memiliki fitur (variabel) dengan skala yang berbeda-beda. Hal ini bisa menyebabkan satu fitur menjadi lebih dominan dibanding fitur lainnya, meskipun tidak lebih penting. Untuk mengatasi hal ini, digunakan teknik standardisasi atau Z-score normalization. Standardisasi mengubah nilai-nilai dalam suatu fitur agar memiliki rata-rata 0 dan standar deviasi 1. Dengan begitu, seluruh fitur berada dalam skala yang sebanding, sehingga model dapat mempelajari pola dengan lebih adil dan akurat [58].

3.6 Feature Selection dan Differentially Expressed Genes

Penelitian ini menerapkan beberapa skenario untuk mengevaluasi pengaruh metode seleksi fitur dan jenis data terhadap performa model klasifikasi stadium kanker payudara. Seleksi fitur dilakukan menggunakan dua pendekatan utama, yaitu:

- **DEG (Differentially Expressed Genes)**: Merupakan fitur yang terseleksi berdasarkan hasil analisis diferensial menggunakan metode *Limma*. Metode ini dipilih karena memiliki performa yang baik dalam menganalisis data yang telah ternormalisasi. Dalam penelitian ini, data ekspresi gen telah melalui proses normalisasi dengan transformasi $\log_2(\text{FPKM-UQ} + 1)$, sehingga metode seperti *DESeq2* atau *edgeR* yang dirancang untuk data mentah atau count-based tidak digunakan. Fitur-fitur yang diperoleh dari hasil analisis DEG kemudian dicocokkan dengan fitur-fitur yang terpilih melalui proses *feature selection*. Hanya fitur yang muncul pada kedua tahap tersebut yang akan digunakan pada proses analisis selanjutnya, sehingga fitur yang digunakan benar-benar mewakili perbedaan signifikan sekaligus memiliki relevansi tinggi terhadap model klasifikasi.

- **Feature Selection (FS)**:
  Proses seleksi fitur dilakukan menggunakan kombinasi antara *Logistic Regression* dengan penalti L1 (Lasso) dan *Recursive Feature Elimination* (RFE). Pada tahap awal, L1 *Logistic Regression* berfungsi untuk menyaring fitur-fitur yang paling berpengaruh secara statistik terhadap label klasifikasi, dengan mengeliminasi fitur yang koefisiennya mendekati nol. Selanjutnya, fitur-fitur yang telah disaring tersebut akan digunakan sebagai *input* dalam proses RFE. Pendekatan berjenjang ini membantu memastikan bahwa fitur yang terpilih benar-benar relevan dan optimal untuk klasifikasi.

Empat skenario diujikan seperti pada Tabel 3.1. Skenario 1 dan 2 menggunakan pendekatan gabungan, yaitu fitur yang dipilih harus termasuk dalam hasil DEG dan juga diseleksi oleh metode RFE dan *L1 Logistic Regression*. Sementara itu, skenario 3 dan 4 hanya menggunakan metode FS tanpa penyaringan awal dengan DEG. Setiap skenario dijalankan pada dua jenis data berbeda, yaitu data ekspresi *miRNA* dan ekspresi gen (FPKM-UQ).

Tabel 3.1. Daftar Skenario Penelitian *Feature Selection* dan *Differentially Expressed Genes*

No.	Skenario	Metode Seleksi Fitur	Jenis Data
1	DEG + FS	DEG (*Limma*) ∩ (RFE + L1 Logistic Regression)	miRNA
2	DEG + FS	DEG (*Limma*) ∩ (RFE + L1 Logistic Regression)	Gene (FPKM-UQ)
3	FS only	RFE + L1 Logistic Regression	miRNA
4	FS only	RFE + L1 Logistic Regression	Gene (FPKM-UQ)

3.7 Pembangunan Model

Pada tahap pembangunan model, fitur-fitur yang telah diseleksi—baik hasil dari kombinasi DEG dan *Feature Selection* (FS), maupun dari FS saja—akan digunakan untuk melatih model klasifikasi. Proses pelatihan dilakukan menggunakan empat skenario algoritma, yaitu *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, dan *Voting Classifier* dengan pendekatan *soft voting*. Pemilihan keempat model ini bertujuan untuk membandingkan performa masing-masing pendekatan dalam mengklasifikasikan stadium kanker payudara.

3.8 Evaluasi

Pada tahap ini, performa model dievaluasi berdasarkan pembagian data menjadi data latih dan data uji dengan rasio 80:20. Evaluasi dilakukan dengan menggunakan sejumlah metrik, yaitu *accuracy*, *precision*, *recall*, dan *f1-score*, yang masing-masing dihitung untuk dua kelas: *early stage* dan *advanced stage*. Selain itu, waktu komputasi untuk setiap proses seleksi dan pelatihan model pada berbagai jumlah fitur juga dicatat sebagai bagian dari pertimbangan efisiensi. Evaluasi ini bertujuan untuk memastikan bahwa model tidak hanya akurat, tetapi juga efisien dan layak diterapkan dalam konteks nyata.

3.9 Analisis Biomarker

Identifikasi biomarker yang signifikan merupakan langkah krusial dalam pengembangan model klasifikasi yang tidak hanya akurat, tetapi juga dapat diterapkan secara efektif dalam praktik klinis. Biomarker yang tepat dapat membantu dalam diagnosis dini, pemantauan perkembangan penyakit, serta

pemilihan strategi pengobatan yang lebih personal. Dalam penelitian ini, kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under the Curve* (AUC) digunakan untuk mengevaluasi kemampuan model dalam membedakan kedua kelas secara menyeluruh. Semakin tinggi nilai AUC, maka semakin baik kemampuan model dalam melakukan klasifikasi stadium kanker payudara secara umum [59]. Selain itu, pencarian kombinasi biomarker potensial dilakukan dengan mengevaluasi performa model pada berbagai jumlah fitur yang diperoleh dari proses seleksi fitur. Tujuan dari pendekatan ini adalah untuk mengidentifikasi kombinasi fitur yang minimal namun tetap mampu memberikan performa klasifikasi yang optimal. Dengan demikian, proses seleksi biomarker tidak hanya mengedepankan tingkat akurasi, tetapi juga mempertimbangkan efisiensi serta potensi aplikasinya dalam praktik klinis, terutama dalam konteks deteksi dan penanganan kanker payudara secara lebih efektif dan terjangkau.