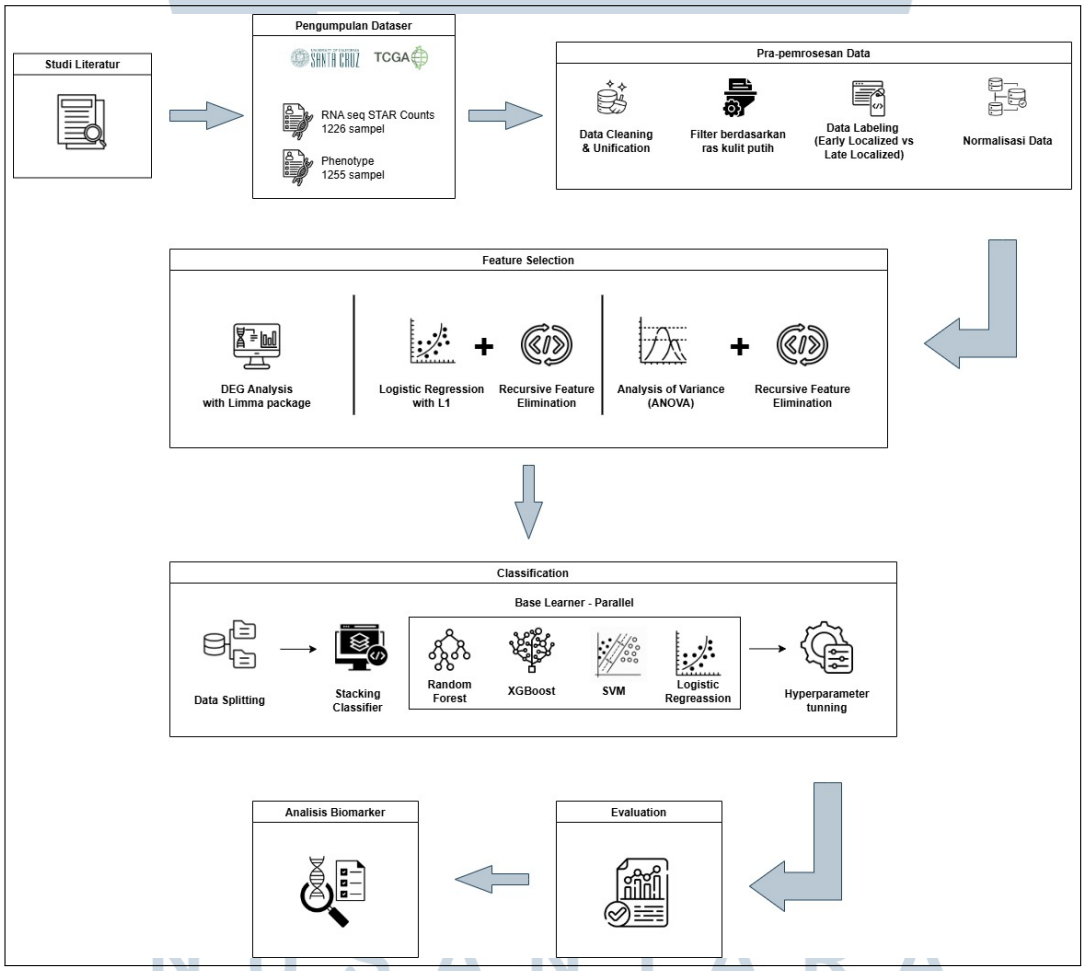
BAB 3
METODOLOGI PENELITIAN

3.1 Alur Kerja Penelitian

Alur kerja penelitian merupakan rangkaian tahapan sistematis yang dijalankan selama proses penelitian. Penyusunan alur ini bertujuan untuk memberikan gambaran menyeluruh mengenai setiap langkah yang dilakukan, mulai dari tahap awal hingga akhir, agar proses penelitian dapat dipahami secara terstruktur, logis, dan ilmiah. Untuk memperjelas, visualisasi alur kerja penelitian disajikan dalam bentuk ilustrasi pada Gambar 3.1.



Gambar 3.1. Alur Penelitian

Dalam penelitian ini, alur kerja mencakup telaah literatur, pengumpulan data, pra-pemrosesan data, seleksi fitur, pembangunan model, serta evaluasi model. Seluruh tahapan tersebut dirancang untuk mencapai hasil yang diinginkan pada penelitian ini.

3.2 Perangkat Penunjang

Selama proses penelitian, menggunakan beberapa perangkat guna mendukung proses pengerjaan penelitian. Berikut adalah beberapa perangkat yang digunakan untuk mendukung proses penelitian:

1. Sebuah PC/Laptop pribadi (ASUS Zeenbook tipe UM462D), dengan spesifikasi lengkap sebagai berikut:
    a. *Operating system*/OS: Windows 10x64bit
    b. CPU: AMD Ryzen 5 - 3500U/BGA
    c. RAM: 8GB
    d. Momory: 512GB-SSD
2. Jupyter Notebook 6.4.12
3. Python 3.9.13
4. Kaggle Notebook (*software web base*), yang memiliki spesifikasi lengkap sebagai berikut:
    a. Auto-saved disk 20GB
    b. CPU: 4 CPU cores, 30 GB RAM
    c. GPU: 1 Nvidia Tesla P100 GPU, 4 CPU cores, 29 GB RAM

3.3 Pengumpulan Data

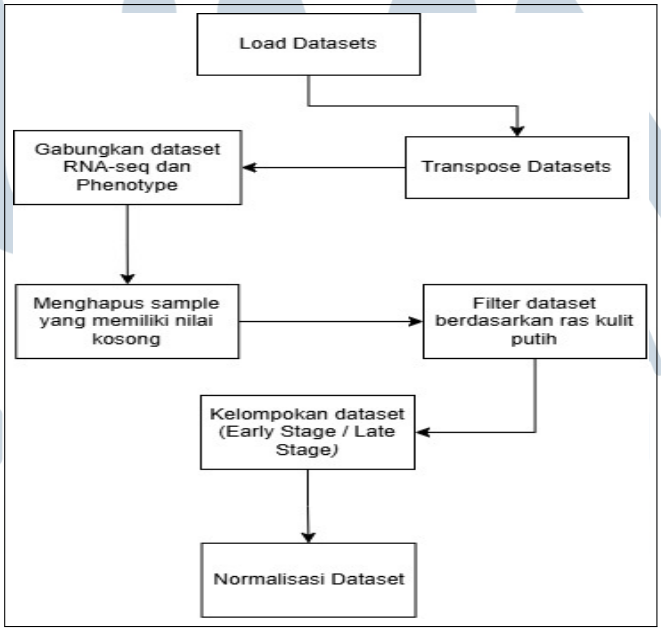
Dataset yang digunakan pada penelitian merupakan dataset dari The Cancer Genome Atlas - Breast Invasive Carcinoma (TCGA BRCA). Dataset ini dapat didapatkan pada website UCSC Xena[34]. Terdapat 2 data yang digunakan pada penelitian, yaitu data ekspresi gen RNA-seq dan data fenotipe (data klinis pasien) dari GDC TCGA Breast Cancer (BRCA).

1. Dataset RNA-seq STAR Counts merupakan data ekspresi gen yang diperoleh dari RNA sequencing yang mengukur ekspresi gen pada tingkat gen menggunakan STAR sebagai hitungan baca mentah (raw read counts). STAR-Counts merupakan sebuah pipeline yang digunakan untuk mengkuantifikasi ekspresi gen dari data RNA-Seq. Proses kuantifikasi dilakukan menggunakan STAR berdasarkan hasil penyelarasan dari sampel tumor atau normal, dan menghasilkan data profil transkriptom. Dataset didownload pada tanggal 05 Maret 2025 dan terdiri dari 60,661 *identifiers* X 1226 *samples*.

2. Dataset fenotipe (*phenotype*) merupakan data klinis dari pasien kanker payudara, data berisi informasi seperti umur, ras, level stadium kanker, dan lain-lain. Dataset didownload pada tanggal 05 Maret 2025 dan terdiri dari 1255 *samples* X 85 *identifiers*.

3.4 Pra-pemrosesan Data (Data Preprocessing)

Pra-pemrosesan data dilakukan untuk memastikan data bersih dan dan konsisten, serta sesuai dengan tujuan penelitian. Terdapat beberapa langkah untuk *data preprocessing* dalam penelitian, seluruh tahap pra-pemrosesan data dilakukan dengan menggunakan Jupyter Notebook dengan basis bahasa pemrograman Python, dapat dilihat pada Gambar 3.2 untuk tahap-tahapnya.



Gambar 3.2. Rangkaian proses pra-proses data

3.4.1 Memuat Dataset

Langkah pertama adalah memuat dataset yang akan digunakan, yaitu *gene expression RNA-seq* dan *phenotype*. Sebagai catatan, data phenotype yang diambil hanya sampel, *stage level*, dan *race* Pemuatan data dilakukan dengan bantuan *package* pandas.

3.4.2 Gabungkan dataset

Langkah selanjutnya adalah menggabungkan dataset RNA-seq dan *phenotype* menjadi satu kesatuan data dengan metode *inner join*. Penggabungan dilakukan berdasarkan nilai yang sama pada kolom sample pada kedua dataset. Penggabungan kedua dataset ini menghasilkan dataset berukuran 1226 *rows* × 60663 *columns*. Setiap baris pada dataset memiliki informasi gene dan informasi *phenotype*.

3.4.3 Menghapus Nilai Kosong (NaN Value)

Langkah berikutnya adalah menghapus sampel yang memiliki nilai NaN sehingga dapat menjaga konsistensi dan integritas dataset. Pada penelitian ini adalah pada nilai-nilai dikolom ras dan level stadium kanker. Setelah menghapus sampel dengan nilai NaN, jumlah sample dalam dataset berkurang dari 1226 menjadi 1213 sampel.

3.4.4 Filter dataset berdasarkan Ras Kulit Putih (White Race)

Langkah selanjutnya adalah melakukan filtrasi berdasarkan ras pasien (kolom *race demographic*). Pada penelitian ini hanya akan mengambil sampel dengan tipe ras kulit putih *non-hispanic* yang merupakan kelompok data dominan dalam dataset, sekitar 71,5% data yang memiliki tipe ras kulit putih. Kemudian, hal ini juga bertujuan untuk menghindari bias karena ekspresi gen dan progresi kanker payudara dapat bervariasi antar kelompok ras/etnis, sehingga dataset konsisten serta stabil. Setelah dilakukan panyaringan ini, dataset berkurang dari 1213 menjadi 867 sampel data.

3.4.5 Filter dataset berdasarkan stadium

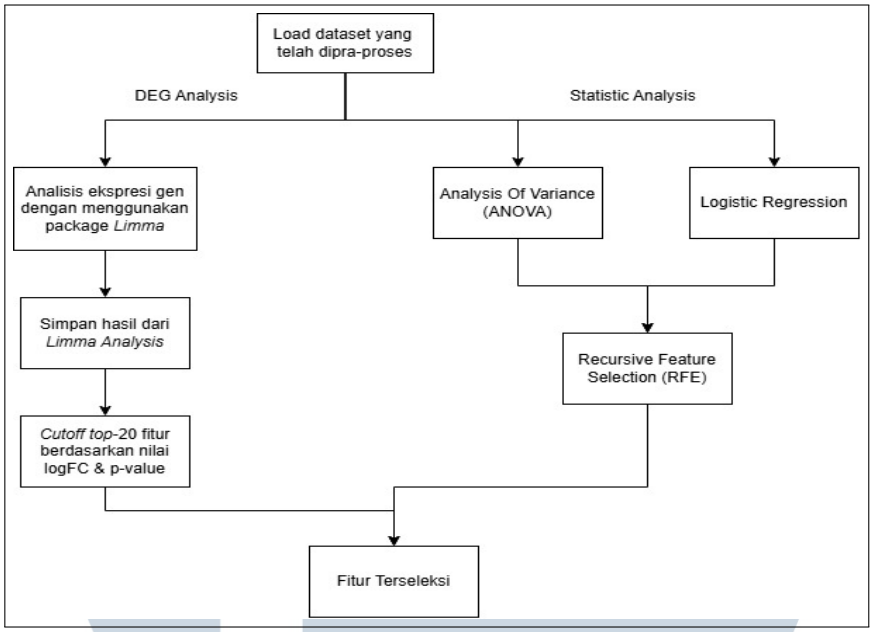
Setelah melakukan filter berdasarkan ras kulit putih, maka selanjutnya melakukan filter dataset berdasarkan stadium kanker payudara. Pada penelitian, dataset yang akan digunakan adalah dataset dengan label kanker stadium II (II, IIA, dan IIB) dan stadium III (III, IIIA, IIIB, dan IIIC). Dataset yang sebelumnya berjumlah 867 sampel, berkurang menjadi 689 sampel dengan pemetaan 485 sampel merupakan data dengan level stadium II dan 204 sampel merupakan data dengan level stadium III. Perbandingan jumlah sampel antara data stadium II dan stadium III adalah 2,38:1.

3.4.6 Labeling Dataset

Tahap selanjutnya adalah melabelkan dataset berdasarkan stadium kanker yang sebelumnya sudah difilter. Akan dibuat kolom baru untuk menandakan *stage level* yang berisi nilai biner/*binary* (0 atau 1). Dataset dengan informasi stadium II (II, IIA, IIB) akan dikategorikan dengan label "**0**" yang menandakan kelompok *Locally advance cancer, early stages*. Kemudian Dataset dengan informasi stadium III (III, IIIA, IIIB, IIIC) akan dikategorikan dengan label "**1**" yang menandakan kelompok *Locally advance cancer, late stages*.

3.5 Seleksi Fitur

Tahapan seleksi fitur dilakukan untuk memilih fitur-fitur yang paling relevan pada sebuah dataset dan membuang fitur yang kurang informatif. Pada penelitian ini, dilakukan beberapa skenario untuk memilih fitur terbaik yang akan digunakan untuk membangun model. Gambar 3.3 merupakan rangkaian proses fitur seleksi.



Gambar 3.3. Rangkaian proses seleksi fitur

Skenario pertama menggunakan metode Differentially Expressed Genes (DEG) yang diidentifikasi melalui pendekatan *limma* untuk mengidentifikasi gen-gen yang memiliki perbedaan ekspresi yang signifikan berbeda antara dua kelas stadium kanker. Data yang telah melalui tahap prapemrosesan akan diproses dengan limma dan akan menghasilkan data ekspresi gen RNA-seq dengan nilai-nilai seperti *Log Fold Change* (logFC), *p-value*, dan *adjusted p-value* (*adj. p-value*). Selanjutnya, melakukan pemilihan fitur-fitur yang dianggap memiliki perbedaan ekspresi yang signifikan dengan memilih *top*-25 *up-regulated* dan *top*-25 *down-regulated* gen yang memiliki nilai p-value < 0.05 dan nilai $|\log \mathrm{FC}| > 0.5$.

Skenario berikutnya menggunakan pendekatan fusion feature selection berbasis statistik ANOVA dan algoritma Recursive Feature Elimination (RFE), yaitu dengan memilih 4000 fitur teratas dari hasil ANOVA atau *Logistic Regresssion* kemudian menyeleksi 50 fitur terbaik menggunakan RFE. Skenario berikutnya juga dilakukan dengan memilih 2000 fitur ANOVA atau *Logistic Regresssion* terlebih dahulu sebelum RFE.

Tabel 3.1. Kombinasi seleksi fitur yang dilakukan

No.	kombinasi seleksi fitue
1	DEG with Limma
2	Logistic Regression 2000 + RFE
3	Logistic Regression 4000 + RFE
4	Anova 2000 + RFE
5	Anova 4000 + RFE

3.6 Pembangunan Model

Setelah proses seleksi fitur, tahap selanjutnya adalah pembangunan model klasifikasi untuk memprediksi stadium kanker payudara berdasarkan ekspresi gen yang telah dipilih. Pada tahap ini, dilakukan pelatihan model klasifikasi menggunakan algoritma *Stacking Classifier*, yaitu pendekatan ensemble learning yang menggabungkan beberapa model pembelajar dasar. Adapun model pembelajar dasar yang digunakan antara lain adalah Random Forest, XGBoost, Support Vector Machine (SVM), dan Logistic Regression.

Tahapan ini dilakukan dengan menggunakan dataset yang telah melalui tahap pra-proses dan seleksi fitur. Kemudian, membagi dataset menjadi data *training* (data latih) dan data *testing* (data uji). Lalu, dilakukan *hyperparameter tuning* pada data *training* untuk mendapatkan kombinasi parameter terbaik untuk model. Selanjutnya, model dengan parameter yang terpilih akan dilatih dengan data *testing* yang dimiliki. Model dari pelatihan data *testing* ini yang akan dilakukan evaluasi performanya.

3.7 Evaluasi Model

Tahap evaluasi model bertujuan untuk mengukur performa model klasifikasi yang telah dilatih sebelumnya. Tahap evaluasi model berfokus pada evaluasi terhadap performa model data *testing* (data uji) yang telah dipisah pada awal klasifikasi.

Evaluasi performa klasifikasi dilakukan dengan pendekatan validasi silang serta pengukuran metrik evaluasi seperti accuracy, precision, recall, F1-score, dan Area Under Curve (AUC-ROC). Pendekatan ensemble stacking ini bertujuan untuk memperoleh model yang lebih stabil dan memiliki generalisasi yang baik dibandingkan dengan model individual.

3.8 Skenario Percobaan

Penelitian ini dilakukan dengan fokus pada klasifikasi biner terhadap stadium II dan stadium III kanker payudara. Selama prosesnya penelitian dilakukan dengan beberapa skenario eksperimen yang telah ditentukan. Skenario-skenario pada penelitian ini menitik beratkan pada metode seleksi fitur yang digunakan. Skenario dalam penelitian dapat dilihat pada tabel 3.2.

Tabel 3.2. Skenario percobaan

No.	Dataset	Kombinasi Seleksi Fitur	Fitur	Model
1	Gen	DEG with Limma	≤ 50	Stacking Classifier
2	Gen	Logistic Regression 2000 + RFE	≤ 50	Stacking Classifier
3	Gen	Logistic Regression 4000 + RFE	≤ 50	Stacking Classifier
4	Gen	Anova 2000 + RFE	≤ 50	Stacking Classifier
5	Gen	Anova 4000 + RFE	≤ 50	Stacking Classifier

3.9 Analisis Biomarker

Analisis biomarker bertujuan untuk mengidentifikasi kandidat biomarker potensial yang memiliki kontribusi penting dalam membedakan stadium II dan stadium III kanker payudara. Setelah seluruh skenario percobaan dievaluasi, dipilih salah satu skenario percobaan yang memberikan kinerja/performa klasifikasi yang terbaik untuk dilakukan analisis biomarker. Analisis dilakukan dengan melakukan studi literatur terhadap gen-gen yang terpilih berdasarkan model terbaik untuk mengidentifikasi keterkaitannya dengan kanker payudara. Gen-gen dari model terbaik tersebut akan dievaluasi kembali dengan pendekatan perhitungan nilai ROC-AUC sebagai indikator evaluasi gen-gen terpilih.