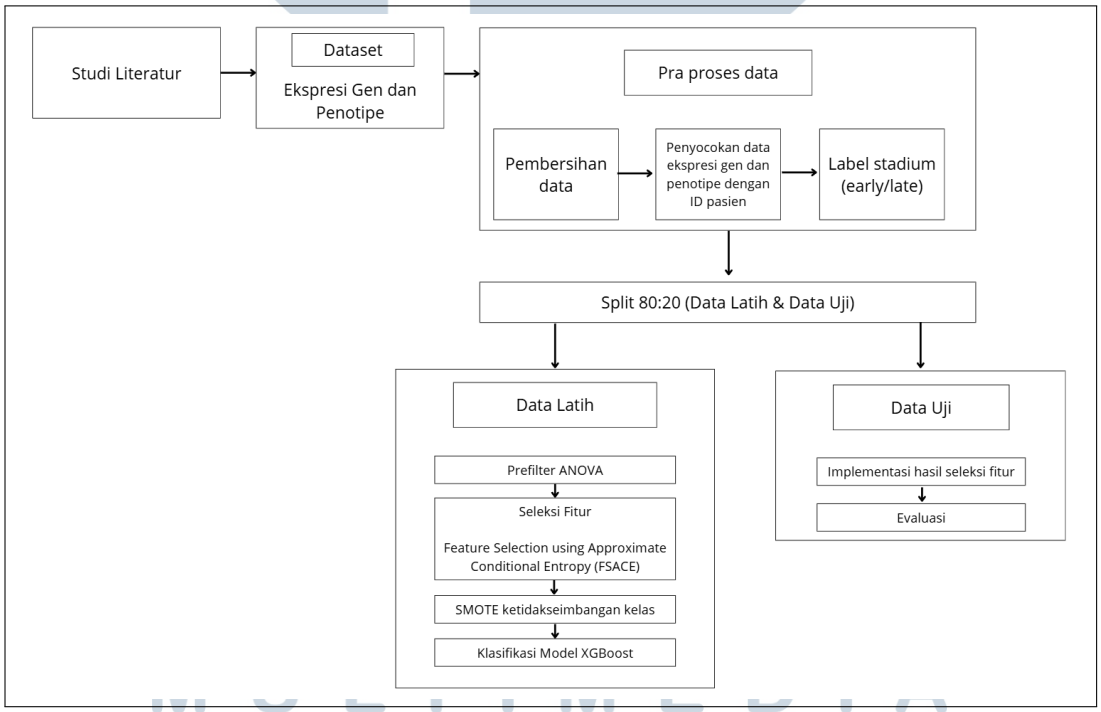
BAB 3
METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian yang digunakan dalam analisis dan klasifikasi kanker usus besar berbasis data ekspresi gen. Pembahasan diawali dengan studi literatur pada bagian 3.1 sebagai landasan teoritis dan metodologis penelitian. Selanjutnya, bagian 3.2 menjelaskan karakteristik dataset yang digunakan, diikuti dengan tahapan praproses data pada bagian 3.3 yang mencakup pelabelan kelas, pemisahan data latih dan data uji, serta prapenyaringan fitur.

Proses seleksi fitur utama menggunakan metode FSACE dibahas secara rinci pada bagian 3.4, yang bertujuan untuk mengatasi permasalahan dimensi tinggi pada data ekspresi gen. Untuk menangani ketidakseimbangan distribusi kelas, diterapkan metode SMOTE sebagaimana dijelaskan pada bagian 3.5. Tahap selanjutnya adalah pembangunan model klasifikasi menggunakan algoritma XGBoost, yang dipaparkan pada bagian 3.6. Evaluasi kinerja model dilakukan menggunakan beberapa metrik pengujian untuk menilai kemampuan generalisasi model dalam membedakan stadium kanker *early* dan *late*. Berikut adalah *pipeline* yang menggambarkan langkah kerja penelitian pada Gambar 3.1.



Gambar 3.1. Pipeline Penelitian

Langkah kerja penelitian akan dijelaskan lebih rinci pada subbab berikutnya, dengan penjabaran setiap tahapan metodologi secara sistematis sesuai dengan alur pada *pipeline* penelitian yang ditunjukkan pada Gambar 3.1.

3.1 Studi Literatur

Untuk mendukung landasan metodologis penelitian ini, dilakukan studi literatur yang berfokus pada beberapa aspek utama, yaitu karakteristik kanker usus besar, pendekatan seleksi fitur pada data ekspresi gen, serta metode klasifikasi biner dengan skema *early* dan *late*. Selain itu, literatur terkait pemodelan klasifikasi pada data berdimensi tinggi (*high dimensional data*) juga dikaji untuk menentukan algoritma pembelajaran mesin yang sesuai. Studi literatur ini digunakan sebagai dasar dalam perancangan alur pra-proses data, pemilihan metode seleksi fitur, serta pembangunan dan evaluasi model klasifikasi dalam penelitian ini.

3.2 Dataset

Penelitian ini menggunakan data ekspresi gen kanker usus besar yang diperoleh dari platform *Xena Browser*, yang mengintegrasikan data genomik dari *The Cancer Genome Atlas* (TCGA). Dataset ekspresi gen yang digunakan merupakan data RNA-Seq dengan format *STAR TPM*, dengan detail sebagai berikut:

1. Data ekspresi gen terdiri dari 515 sampel pasien dengan jumlah fitur sebanyak 60.660 gen untuk setiap sampel yang diambil dari GDC TCGA Colon Cancer https://xenabrowser.net/.
2. Data klinis pasien yang terdiri 562 pasien dan 86 atribut informasi klinis yang diambil dari GDC TCGA Colon Cancer https://xenabrowser.net/.

Dataset ini digunakan sebagai dasar analisis fitur ekspresi gen dan pengujian model klasifikasi. Data ekspresi gen akan diproses untuk seleksi fitur menggunakan FSACE, sedangkan data klinis digunakan sebagai informasi tambahan untuk memvalidasi hubungan *(ground truth)* antara fitur gen dan atribut klinis berdasarkan nomor pasien (TCGA ID).

3.3 Pra-proses data

Setelah dataset diperoleh, tahap selanjutnya adalah menyelaraskan data dengan mencocokkan identitas pasien berdasarkan kode TCGA. Data ekspresi gen dan data klinis yang telah digabungkan kemudian diberi label kelas berdasarkan stadium kanker pasien yang tercantum dalam sistem staging AJCC. Stadium patologis pasien yang semula terdiri dari *Stage* I (termasuk IA, IB), *Stage* II (termasuk IIA, IIB, IIC), *Stage* III (termasuk IIIA, IIIB, IIIC), hingga *Stage* IV (termasuk IVA dan IVB) selanjutnya dikelompokkan menjadi dua kelas utama. Pasien dengan stadium I dan II beserta seluruh variannya digabungkan ke dalam kelas *early*, sedangkan pasien dengan stadium III dan IV beserta seluruh variannya digabungkan ke dalam kelas *late*.

Setelah proses pelabelan dilakukan, dataset kemudian dipisahkan menjadi data latih dan data uji sebelum dilakukannya proses seleksi fitur dengan rasio 80 banding 20. Pemisahan ini bertujuan untuk mencegah terjadinya *data leakage*, yaitu kondisi ketika informasi dari data uji secara tidak langsung memengaruhi proses pemilihan fitur, sehingga dapat menyebabkan estimasi kinerja model terlihat lebih baik dari kondisi sebenarnya.

Selanjutnya, data latih dibersihkan dengan menghapus gen yang memiliki nilai variansi di bawah ambang batas sebesar 0.1, yang bertujuan untuk mengeliminasi gen dengan tingkat variasi

yang sangat rendah. Tahap ini dilakukan untuk mengurangi fitur yang bersifat hampir konstan dan minim informasi. Setelah itu, data latih diproses menggunakan metode ANOVA sebagai tahap prapenyaringan fitur untuk mengidentifikasi gen yang memiliki perbedaan ekspresi yang signifikan antar kelas. Proses ini bertujuan untuk menurunkan dimensi data sebelum dilakukan seleksi fitur lanjutan menggunakan FSACE.

3.4 Seleksi Fitur

Pada penelitian ini, FSACE digunakan sebagai metode seleksi fitur utama untuk mengatasi permasalahan dimensi tinggi pada data ekspresi gen kanker usus besar. Setelah tahap praproses dan prapenyaringan fitur menggunakan *variance threshold* dan ANOVA, FSACE diterapkan pada data latih untuk memilih gen yang paling relevan terhadap kelas stadium kanker.

FSACE bekerja dengan mengevaluasi kontribusi setiap gen berdasarkan penurunan nilai *approximate conditional entropy* (H_{ace}) terhadap atribut keputusan. Hubungan kemiripan antar sampel dibentuk menggunakan relasi *fuzzy* berbasis *Laplacian kernel*, sehingga metode ini mampu menangani data kontinu dan ketidakpastian yang umum terdapat pada data ekspresi gen.

Pada proses awal seleksi fitur, atribut inti (*core attributes*) ditentukan berdasarkan nilai *Importance of Internal Attribute* (IIA). Selanjutnya, fitur kandidat dievaluasi menggunakan *Importance of External Attribute* (IEA), dan fitur dengan nilai IEA tertinggi ditambahkan ke dalam subset fitur. Proses ini dihentikan ketika penurunan nilai H_{ace} tidak lagi signifikan atau jumlah fitur telah mencapai batas yang ditentukan.

Subset gen hasil seleksi FSACE kemudian digunakan sebagai input pada tahap pembangunan model klasifikasi menggunakan XGBoost.

3.5 Solusi Data Tidak Seimbang

Untuk mengatasi permasalahan ketidakseimbangan distribusi kelas pada data latih, diterapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Proses balancing ini hanya dilakukan pada data latih yang telah melalui seleksi fitur, sedangkan data uji tidak mengalami perubahan distribusi kelas. Hal ini bertujuan untuk menjaga objektivitas evaluasi model serta mencegah terjadinya *data leakage*.

3.6 Pembangunan Model

Model klasifikasi kemudian dibangun menggunakan algoritma XGBoost dengan memanfaatkan fitur hasil seleksi FSACE. Parameter model ditentukan secara manual berdasarkan pertimbangan empiris dan eksperimen awal, tanpa menerapkan metode *hyperparameter tuning*. Pendekatan ini dipilih untuk menilai kontribusi seleksi fitur terhadap kinerja model secara lebih objektif.

Model yang telah dilatih selanjutnya diimplementasikan pada data uji dengan menggunakan subset fitur yang sama seperti pada data latih, tanpa melakukan proses seleksi ulang. Tahap ini bertujuan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

Evaluasi kinerja model dilakukan menggunakan beberapa metrik, yaitu akurasi, confusion matrix, serta kurva *Receiver Operating Characteristic* (ROC) dan *Area Under the Curve* (AUC). Hasil evaluasi ini digunakan sebagai dasar analisis performa model klasifikasi kanker usus besar berbasis ekspresi gen.