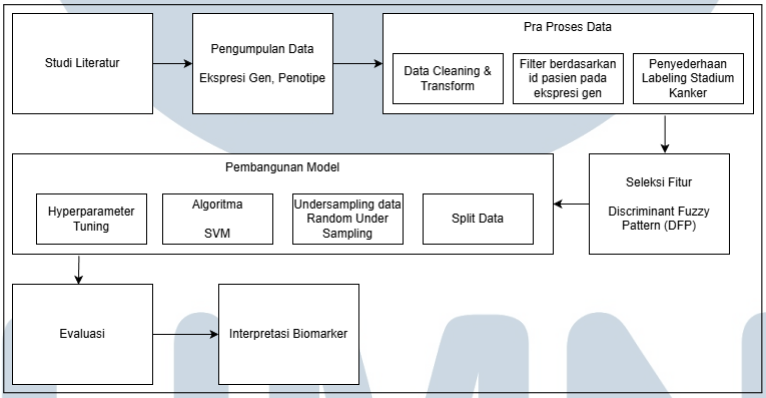
BAB 3
METODOLOGI PENELITIAN

Penelitian ini dirancang untuk memverifikasi efikasi profil ekspresi gen dalam membedakan antara kanker payudara stadium awal dan lanjut melalui kerangka kerja komputasional yang presisi. Untuk mengatasi tantangan data *High-Dimensional Low-Sample Size* (HDLSS) dan ketidakpastian batas kelas yang inheren pada dataset biologis, diterapkan pendekatan hibrida yang mengintegrasikan DFP untuk seleksi fitur dan SVM untuk klasifikasi. Metode DFP diprioritaskan karena kemampuannya memodelkan keanggotaan fitur *fuzzy*, yang memungkinkan pelestarian informasi diskriminatif yang sering kali diabaikan oleh pendekatan statistik konvensional atau metode reduksi berbasis varians (seperti PCA) yang cenderung melewatkan fitur bervarians rendah namun signifikan secara biologis. Selain itu, SVM dipilih dibandingkan model *Deep Learning* yang kompleks untuk memitigasi risiko *overfitting* yang tinggi, mengingat terbatasnya ukuran sampel klinis yang tersedia. Alur kerja sistematis penelitian ini, mulai dari akuisisi data hingga interpretasi *biomarker*, diringkas secara visual pada Gambar 3.1.



Gambar 3.1. Research pipeline untuk data gen

3.1 Studi Literatur

Tahapan ini bertujuan untuk memperoleh landasan teori yang kuat mengenai topik penelitian. Kajian dilakukan terhadap literatur ilmiah terkini yang membahas diagnosis kanker payudara, analisis data multimodal, metode feature selection, serta algoritma klasifikasi berbasis kecerdasan buatan. Studi literatur membantu mengidentifikasi kesenjangan penelitian (research gap) yang menjadi dasar pengembangan model pada penelitian ini, khususnya terkait integrasi data ekspresi gen serta penerapan metode DFP.

3.2 Pengumpulan Data

Tahap pengumpulan data pada penelitian ini dilakukan dengan memanfaatkan repositori publik UCSC Xena Browser, yang menyediakan akses terbuka terhadap berbagai data genomik dan

klinis dari proyek The Cancer Genome Atlas (TCGA). Dataset yang digunakan secara khusus berasal dari Cohort GDC TCGA Breast Cancer (BRCA), yang merupakan kumpulan data ekspresi gen dan penotipe pasien kanker payudara. Deskripsi dataset didefinisikan pada tabel 3.1

Tabel 3.1. Deskripsi dataset yang digunakan

Informasi	**RNA-seq (STAR TPM)**	**Phenotype**
Jumlah sampel	1226	1255
Versi	05-20-2024	09-07-2024
Unit	$log2(\text{tpm}+1)$	—
Tipe data	Ekspresi gen	Data fenotipe
Jumlah kolom	60.661 gen	85 atribut

Data ekspresi gen diperoleh dalam format RNA-Seq – STAR – TPM (Transcripts Per Million) yang telah melalui proses penyelarasan (alignment) menggunakan Spliced Transcripts Alignment to a Reference (STAR) dan normalisasi berbasis TPM, sehingga hasilnya dapat dibandingkan antar sampel secara kuantitatif. Dataset ini mencakup 1.226 pasien kanker payudara, dengan ribuan gen yang diekspresikan pada masing-masing sampel.

Selain data ekspresi gen, penelitian ini juga memanfaatkan data penotipe (phenotype data) yang tersedia dalam repositori yang sama. Data penotipe berisi informasi klinis pasien, termasuk stadium kanker yang diklasifikasikan berdasarkan sistem American Joint Committee on Cancer (AJCC). Data penotipe ini berfungsi sebagai label kelas (ground truth) dalam proses klasifikasi stadium kanker payudara.

3.3 Pra-Proses Data

Tahap pra-proses data dilakukan untuk memastikan data ekspresi gen dan fenotipe memiliki format yang sesuai, bersih, dan konsisten sebelum masuk ke tahap pemodelan. Langkah awal difokuskan pada penyesuaian struktur data ekspresi gen melalui proses *transposisi*. Data mentah yang semula menempatkan gen sebagai baris dan sampel sebagai kolom diubah orientasinya, sehingga setiap baris merepresentasikan satu sampel pasien dan setiap kolom merepresentasikan fitur genetik. Hal ini dilakukan untuk memenuhi standar input algoritma *machine learning*.

Selanjutnya, dilakukan pembersihan pada data fenotipe klinis. Mengingat variasi keterisian data antar atribut, dilakukan eliminasi terhadap fitur (kolom) yang memiliki nilai kosong (*null*) melebihi 50% dari total populasi sampel. Langkah ini bertujuan untuk membuang atribut yang informasinya terlalu sedikit (*sparse*) dan mempertahankan fitur yang representatif.

Setelah data dibersihkan, dilakukan proses kategorisasi dan pelabelan ulang (*re-labeling*) pada kolom stadium kanker (*pathologic stage*). Kategori stadium awal yang beragam disederhanakan menjadi skenario klasifikasi biner. Sampel dengan label Stage I dan Stage II dikelompokkan menjadi kelas *Early Stage* (diberi label 0), sedangkan sampel dengan label Stage III dikelompokkan menjadi kelas *Late Stage* (diberi label 1).

Tahap berikutnya adalah penggabungan data (*data merging*) antara data ekspresi gen dan label fenotipe yang telah dikategorikan. Integrasi dilakukan menggunakan metode irisan (*inner join*) berdasarkan ID Sampel pasien. Hanya pasien yang memiliki data lengkap pada kedua dataset (ekspresi gen dan fenotipe) yang dipertahankan, sedangkan sampel yang tidak berpasangan dieksklusi untuk menjaga integritas analisis.

Langkah terakhir dalam pra-proses adalah normalisasi data ekspresi gen menggunakan teknik standardisasi (*Z-Score Normalization*). Setiap nilai fitur genetik ditransformasi sehingga memiliki rata-rata (*mean*) 0 dan simpangan baku (*standard deviation*) 1. Proses ini krusial untuk menyamakan skala antar fitur gen yang memiliki rentang nilai ekspresi berbeda, sehingga mencegah bias pada perhitungan jarak dalam algoritma SVM.

3.4 Seleksi Fitur

Data ekspresi gen memiliki jumlah fitur yang sangat besar (ribuan gen) dengan jumlah sampel relatif sedikit, yang dikenal sebagai permasalahan *high-dimensional low-sample size* (HDLSS). Untuk mengatasi hal ini, dilakukan seleksi fitur menggunakan metode DFP. DFP merupakan pendekatan berbasis fuzzy logic yang mengukur kemampuan diskriminatif setiap fitur berdasarkan derajat keanggotaan (fuzzy membership function) terhadap kelas tertentu. Melalui konsep ini, setiap gen dievaluasi berdasarkan seberapa besar kontribusinya dalam membedakan antar kelas kanker. Fitur dengan nilai diskriminatif tertinggi dipertahankan, sedangkan fitur dengan kontribusi rendah dieliminasi. Pendekatan ini tidak hanya mengurangi dimensi data, tetapi juga mempertahankan informasi biologis penting yang berpotensi relevan terhadap perkembangan kanker payudara.

Mengingat kinerja metode DFP sangat dipengaruhi oleh penentuan nilai ambang batas (*threshold*) dalam pembentukan pola fuzzy, penggunaan parameter default tidak selalu menjamin hasil yang optimal pada setiap karakteristik dataset. Oleh karena itu, penelitian ini menerapkan mekanisme hyperparameter tuning menggunakan metode Grid Search untuk mengeksplorasi ruang parameter yang paling efektif dalam memisahkan kelas kanker. *Hyperparameter tuning* dilakukan dengan menguji berbagai kombinasi nilai pada tiga parameter utama DFP, yaitu Skip Factor, Pi Value, dan Zeta. Rincian konfigurasi parameter yang diujikan dalam penelitian ini dirangkum pada Tabel 3.2.

Tabel 3.2. Hyperparameter tuning DFP

Model	**Parameter**	**Nilai**
Discriminant Fuzzy Pattern	Skip Factor	1, 1.5, 2, 2.5, 3
	Pi Value	0.5, 0.6, 0.65, 0.7, 0.75
	Zeta (ζ)	0.1, 0.2, 0.3, 0.4, 0.5

Variasi nilai pada ketiga parameter di atas berperan krusial dalam mengendalikan sensitivitas dan selektivitas algoritma terhadap fitur-fitur yang dianggap relevan. Parameter Pi Value dan Zeta (ζ) berfungsi sebagai pengatur batas toleransi ketidakpastian (fuzziness) dan derajat keanggotaan, yang secara langsung menentukan seberapa ketat kriteria seleksi gen dilakukan.

Sementara itu, Skip Factor digunakan untuk menyesuaikan granularitas dalam pencarian pola data. Dengan mengevaluasi kombinasi nilai-nilai tersebut, tujuan utamanya adalah mendapatkan konfigurasi optimal yang mampu mereduksi dimensi secara signifikan namun tetap mempertahankan akurasi klasifikasi tertinggi.

3.5 Pembangunan Model

Setelah himpunan fitur genetik yang paling diskriminatif diperoleh melalui seleksi fitur DFP, tahapan selanjutnya berfokus pada pembangunan model klasifikasi prediktif menggunakan algoritma SVM. Langkah awal dalam fase ini adalah partisi data (data splitting) untuk memisahkan dataset menjadi data latih (training set) dan data uji (testing set) dengan rasio 80:20 dan 90:10. Proses pembagian ini dilakukan menggunakan teknik stratified sampling, yang bertujuan untuk menjaga proporsi kelas stadium awal dan lanjut agar tetap konsisten di kedua subset, serta mencegah terjadinya kebocoran informasi (data leakage) yang dapat membiaskan evaluasi model.

Mengingat tantangan ketidakseimbangan distribusi kelas (*class imbalance*) yang ditemukan pada data ekspresi gen, penelitian ini menerapkan strategi penyeimbangan data pada data latih menggunakan metode *Random Under Sampling* (RUS). Berbeda dengan teknik *oversampling* yang membangkitkan data sintetis, RUS bekerja dengan mengurangi jumlah sampel pada kelas mayoritas (*Early Stage*) secara acak hingga proporsinya setara dengan kelas minoritas (*Late Stage*). Langkah pra-pemrosesan ini dilakukan untuk mencegah bias model terhadap kelas mayoritas, sehingga objektivitas klasifikasi dapat terjaga tanpa mengubah integritas data uji asli.

Untuk menjamin performa model yang optimal, dilakukan optimasi hiperparameter (hyperparameter tuning) menggunakan metode Grid Search Cross-Validation (GridSearchCV). Metode ini secara sistematis mengevaluasi kombinasi parameter SVM—yaitu parameter regularisasi (C), jenis kernel (Linear dan RBF), serta koefisien kernel (γ)—untuk menemukan konfigurasi yang menghasilkan skor validasi terbaik. Daftar parameter yang dioptimasi disajikan pada Tabel 3.3.

Tabel 3.3. Parameter model SVM

Model	**Parameter**	**Nilai**
Support Vector Machine	C	0.1, 1, 10, 100
	kernel	linear, rbf
	γ	1, 0.1, 0.01, 0.001

Sebagai langkah penyempurnaan akhir, penelitian ini tidak menggunakan ambang batas (threshold) probabilitas standar 0.5 untuk klasifikasi. Sebaliknya, dilakukan Optimasi Decision Threshold berbasis kurva Precision-Recall. Nilai ambang batas digeser secara dinamis untuk mengidentifikasi titik potong (cutoff point) yang memaksimalkan nilai F1-Score. Pendekatan ini memastikan bahwa model akhir memiliki keseimbangan terbaik antara presisi dan recall dalam mendeteksi stadium kanker, sehingga prediksi yang dihasilkan lebih akurat dan dapat diandalkan secara klinis.

3.6 Evaluasi Model

Kinerja model dievaluasi menggunakan berbagai metrik yang diperoleh dari Confusion Matrix, antara lain Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, serta Area Under the Curve (AUC) dari kurva ROC. Masing-masing metrik memberikan perspektif berbeda terhadap performa model, seperti ketepatan prediksi keseluruhan, kemampuan mendeteksi kasus positif, serta keseimbangan antara sensitivitas dan presisi. ROC Curve digunakan untuk menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai ambang batas, sementara AUC digunakan untuk mengukur kemampuan diskriminatif model secara keseluruhan.

3.7 Interpretasi Biomarker

Tahap akhir penelitian ini adalah interpretasi biomarker, yaitu analisis terhadap gen-gen yang terpilih melalui metode DFP untuk mengidentifikasi potensi biomarker yang berperan penting dalam perkembangan kanker payudara. Gen dengan bobot kontribusi tinggi dievaluasi berdasarkan literatur biologis dan basis data molekuler untuk memahami fungsinya dalam proses biologis seperti proliferasi sel, apoptosis, angiogenesis, atau metastasis. Tahap ini memberikan konteks biologis terhadap hasil klasifikasi dan membantu mengaitkan temuan komputasional dengan mekanisme nyata dalam patologi kanker payudara, sehingga meningkatkan validitas dan potensi aplikasi klinis dari hasil penelitian ini.