BAB 2
LANDASAN TEORI

2.1 Penelitian Terdahulu

Sejumlah penelitian terdahulu telah menerapkan machine learning pada data ekspresi gen untuk klasifikasi kanker, baik secara umum maupun khusus pada kanker prostat. Bagian ini membahas penelitian-penelitian tersebut beserta metode, data, hasil, dan kaitannya dengan penelitian ini.

Tinjauan terhadap penelitian terdahulu memiliki dua tujuan utama. Pertama, untuk memetakan pendekatan yang telah digunakan sehingga posisi penelitian ini terhadap pengetahuan yang sudah ada menjadi jelas. Kedua, untuk mengidentifikasi kekurangan atau celah yang belum tertangani, yang kemudian menjadi landasan bagi kontribusi penelitian ini. Penelitian-penelitian yang ditinjau dipilih karena relevan dari sisi topik, yaitu klasifikasi kanker berbasis ekspresi gen, maupun dari sisi metode, yaitu penerapan berbagai algoritma machine learning mulai dari model sederhana hingga model kompleks seperti deep learning dan ensemble. Dengan menelaah kekuatan dan keterbatasan masing-masing, penelitian ini dapat mengambil pelajaran yang baik sekaligus menghindari kelemahan yang berulang, khususnya terkait dengan ketatnya protokol evaluasi yang sering kurang diperhatikan.

Alharbi dan Vakanski [9] melakukan tinjauan terhadap berbagai metode machine learning untuk klasifikasi kanker menggunakan data ekspresi gen. Tinjauan tersebut menyimpulkan bahwa ekspresi gen berperan penting dalam deteksi dini kanker karena mencerminkan proses biokimia pada sel dan jaringan, namun tingginya dimensi data menjadi tantangan utama yang menuntut seleksi fitur yang tepat. Penelitian ini sejalan dengan temuan tersebut dengan menggunakan seleksi gen berbasis DESeq2 untuk mereduksi 60.488 gen menjadi tujuh gen yang paling diskriminatif.

Haddou Bouazza [10] mengevaluasi beberapa model machine learning untuk klasifikasi kanker prostat menggunakan profil ekspresi gen dari DNA microarray, dengan menerapkan berbagai metode seleksi fitur seperti Signal to Noise Ratio, ReliefF, dan Correlation Coefficient untuk menghilangkan gen yang redundan, dengan kombinasi terbaik (Signal to Noise Ratio dan Linear Discriminant Analysis) mencapai akurasi klasifikasi 95% [10]. Penelitian tersebut menegaskan pentingnya reduksi dimensi, namun seleksi fitur dilakukan terpisah

dari skema validasi, sehingga berpotensi menimbulkan kebocoran data. Penelitian ini mengatasi celah tersebut dengan melakukan seleksi gen di dalam setiap lipatan pelatihan.

Marouf dkk. [11] mengidentifikasi biomarker potensial pada kanker prostat menggunakan classifier machine learning yang dapat dijelaskan (explainable), dengan akurasi tertinggi 81,01% menggunakan Random Forest [11]. Penelitian tersebut menekankan pentingnya interpretabilitas dalam penemuan biomarker. Penelitian ini melengkapi pendekatan tersebut dengan menggunakan Regresi Logistik yang transparan serta memvalidasi gen yang ditemukan terhadap literatur biologis.

Pada ranah deep learning, Alici-Karaca dan Akay [13] serta Horasan dan Güneş [14] mengembangkan model deep learning untuk diagnosis kanker prostat dengan integrasi citra Magnetic Resonance Imaging. Horasan dan Güneş [14] secara khusus menggunakan beberapa arsitektur deep learning, yaitu 3D Convolutional Neural Network, Residual Network, dan Inception Network, untuk meningkatkan akurasi dan ketahanan deteksi. Pendekatan deep learning memberikan akurasi tinggi tetapi membutuhkan data besar dan menghasilkan model yang kurang transparan. Penelitian ini memilih model sederhana yang interpretable karena jumlah fitur yang sedikit (tujuh gen) tidak memerlukan kompleksitas deep learning.

Mahmoud dan Takaoka [20] mengembangkan stacking ensemble dengan seleksi fitur untuk klasifikasi kanker berbasis ekspresi gen dan melaporkan peningkatan akurasi. Penelitian ini menggunakan stacking ensemble (XGBoost, Random Forest, LightGBM) sebagai pembanding, namun menemukan bahwa model sederhana Regresi Logistik tidak kalah, sehingga model sederhana dipilih sesuai prinsip parsimoni. Selain itu, Mahin dkk. [15] dan Rao serta Prasad [16] menunjukkan efektivitas machine learning untuk klasifikasi pan-kanker dan data RNA-Seq, yang memperkuat relevansi pendekatan berbasis ekspresi gen.

Ringkasan penelitian terdahulu disajikan pada Tabel 2.1. Secara umum, penelitian-penelitian tersebut menunjukkan bahwa fitur berbasis ekspresi gen efektif untuk membedakan jaringan kanker dan normal. Namun, sebagian besar belum secara eksplisit menangani isu kebocoran data pada tahap seleksi fitur maupun pembagian per-pasien, sehingga estimasi performa berpotensi terlalu optimis. Celah penelitian (research gap) inilah yang menjadi fokus penelitian ini, yaitu menyajikan tanda tangan gen yang stabil dan estimasi performa yang bebas kebocoran serta dapat direproduksi.

Tabel 2.1. Ringkasan Penelitian Terdahulu

Penulis (Tahun)	**Metode**	**Domain**	**Temuan Utama**
Alharbi & Vakanski (2023) [9]	Tinjauan ML	Klasifikasi kanker	Mengulas metode ML pada ekspresi gen untuk kanker
Haddou Bouazza (2024) [10]	Beberapa model ML	Kanker prostat	Evaluasi model ML pada profil ekspresi gen prostat
Marouf dkk. (2025) [11]	Explainable ML	Kanker prostat	Identifikasi biomarker dengan klasifikasi yang dapat dijelaskan
Alici-Karaca & Akay (2024) [13]	Deep learning	Kanker prostat	Model deep learning efisien untuk diagnosis
Mahmoud & Takaoka (2025) [20]	Stacking ensemble	Kanker (ekspresi gen)	Ensemble dengan seleksi fitur meningkatkan akurasi

2.1.1 Celah Penelitian

Berdasarkan tinjauan penelitian terdahulu di atas, dapat dibentuk celah penelitian (research gap) yang menjadi dasar penelitian ini. Sebagian besar penelitian terdahulu berfokus pada peningkatan akurasi klasifikasi, tetapi belum secara eksplisit menyatakan penanganan kebocoran data pada dua titik kritis, yaitu seleksi fitur dan pembagian data. Seleksi fitur yang dilakukan pada seluruh data sebelum validasi silang, serta pembagian data yang tidak dikelompokkan per-pasien, dapat menghasilkan estimasi performa yang terlalu optimis. Selain itu, sebagian penelitian menggunakan model yang kurang transparan sehingga sulit diinterpretasikan, dan tidak semua memvalidasi gen temuan terhadap literatur biologis. Perbandingan penelitian ini terhadap penelitian terdahulu pada keempat aspek tersebut disajikan pada Tabel 2.2. Penelitian ini mengisi celah tersebut dengan menerapkan seleksi gen di dalam lipatan, pembagian per-pasien, model

yang transparan, sekaligus validasi biologis terhadap gen temuan.

Tabel 2.2. Posisi Penelitian Ini terhadap Penelitian Terdahulu

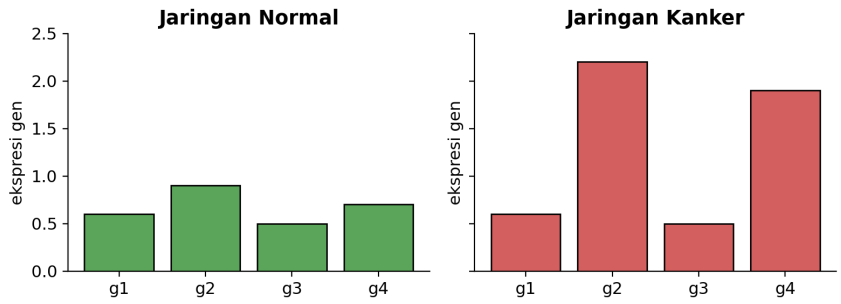
Penelitian	**Seleksi in-fold**	**Split per-pasien**	**Model transparan**	**Validasi biologis**
Haddou Bouazza (2024)	Tidak	Tidak dinyatakan	Tidak dinyatakan	Tidak dinyatakan
Marouf dkk. (2025)	Tidak dinyatakan	Tidak dinyatakan	Ya	Tidak dinyatakan
Alici-Karaca & Akay (2024)	Tidak dinyatakan	Tidak dinyatakan	Tidak	Tidak dinyatakan
Mahmoud & Takaoka (2025)	Tidak dinyatakan	Tidak dinyatakan	Tidak	Tidak dinyatakan
Penelitian ini	**Ya**	**Ya**	**Ya**	**Ya**

2.2 Kanker Prostat

Kanker prostat adalah keganasan yang berasal dari sel-sel kelenjar prostat. Pada tingkat molekuler, kanker ditandai oleh perubahan pola ekspresi gen yang menyimpang dari jaringan normal. Deteksi dini sangat menentukan prognosis, sehingga penemuan biomarker molekuler yang andal menjadi penting untuk melengkapi metode diagnosis konvensional [5]. Perbedaan pola ekspresi gen antara jaringan normal dan jaringan kanker diilustrasikan pada Gambar 2.1.

Secara biologis, perkembangan kanker prostat merupakan proses bertahap yang melibatkan akumulasi perubahan genetik dan epigenetik pada sel epitel kelenjar prostat. Sel yang semula tumbuh terkendali kehilangan mekanisme pengaturan siklus sel sehingga berproliferasi secara berlebihan. Perubahan ini tercermin pada tingkat transkripsi, yaitu sejumlah gen menjadi terekspresi berlebih (upregulated) sementara gen lain tertekan (downregulated). Karena itu, pola ekspresi gen dapat berfungsi sebagai ”sidik jari” molekuler yang membedakan jaringan kanker dari jaringan normal. Tingkat keganasan kanker prostat secara klinis dinilai menggunakan skor Gleason, yang mengelompokkan pola pertumbuhan sel dari yang paling terdiferensiasi hingga paling agresif. Kanker dengan skor Gleason tinggi (7–10) cenderung tumbuh cepat dan bermetastasis, sedangkan skor rendah sering bersifat indolen. Ketidakmampuan metode konvensional untuk membedakan kedua kelompok ini secara pasti menjadi salah satu alasan pentingnya penanda molekuler berbasis ekspresi gen, yang berpotensi menangkap

perbedaan pada tataran yang lebih mendasar dibandingkan penilaian morfologi semata.



Gambar 2.1. Perbedaan Pola Ekspresi Gen Kanker dan Normal

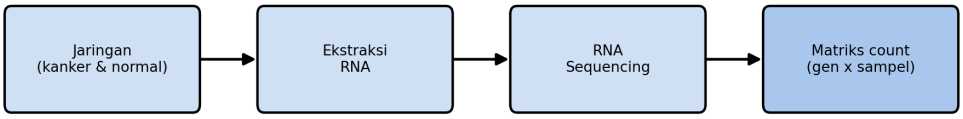
Secara biologis, prostat adalah kelenjar berukuran kecil pada sistem reproduksi pria yang berfungsi menghasilkan sebagian cairan semen. Kanker prostat berkembang ketika sel-sel epitel kelenjar prostat mengalami akumulasi mutasi dan disregulasi ekspresi gen sehingga tumbuh serta membelah secara tidak terkendali. Tingkat keganasan kanker prostat secara klinis dinilai menggunakan skor Gleason, yang mengevaluasi derajat diferensiasi sel pada jaringan biopsi. Skor Gleason rendah (6 atau kurang) umumnya menunjukkan kanker yang indolen dan tumbuh lambat, sedangkan skor tinggi (7 hingga 10) berkaitan dengan subtipe agresif yang berisiko bermetastasis ke tulang dan kelenjar getah bening. Heterogenitas inilah yang membuat penanganan kanker prostat menjadi kompleks, karena tidak semua kasus memerlukan intervensi agresif, sementara sebagian kecil kasus justru berkembang dengan cepat dan mematikan [5].

Pada tingkat molekuler, transformasi keganasan pada kanker prostat kerap melibatkan jalur pensinyalan reseptor androgen (androgen receptor signaling) yang berperan sentral dalam pertumbuhan sel prostat. Selain itu, terjadi pula perubahan pada gen-gen yang mengatur metabolisme lipid, proliferasi sel, serta interaksi dengan matriks ekstraseluler [34, 35]. Perubahan pola ekspresi gen ini bersifat sistematis dan dapat diukur, sehingga membuka peluang untuk menemukan tanda tangan molekuler (molecular signature) berupa sekelompok gen yang secara konsisten menyimpang ekspresinya pada jaringan kanker. Tanda tangan semacam ini berpotensi menjadi penanda objektif yang melengkapi penilaian histopatologi konvensional yang bersifat subjektif [5]. Oleh karena itu, pendekatan berbasis data ekspresi gen menjadi relevan untuk menemukan penanda yang tidak hanya akurat, tetapi juga dapat ditelusuri perannya secara biologis [6].

2.3 Data Ekspresi Gen RNA-Seq dan TCGA

RNA-Sequencing mengukur jumlah transkrip (count) tiap gen pada suatu sampel. Data TCGA-PRAD menyediakan matriks ekspresi 60.488 gen untuk ratusan sampel jaringan kanker dan normal beserta metadata klinis. Data semacam ini banyak dimanfaatkan dalam analisis bioinformatika untuk menemukan gen penanda dan membangun model prediktif [6, 12]. Alur ringkas dari jaringan hingga matriks ekspresi ditunjukkan pada Gambar 2.2.

Proses RNA-Sequencing diawali dengan ekstraksi RNA dari jaringan, konversi menjadi complementary DNA (cDNA), fragmentasi, lalu pembacaan sekuens oleh mesin sequencing. Fragmen bacaan (reads) kemudian dipetakan ke genom referensi, dan jumlah bacaan yang jatuh pada tiap gen dihitung sebagai nilai count. Nilai count inilah yang menjadi representasi tingkat ekspresi gen. Karena setiap sampel dapat memiliki kedalaman sekuensing (total bacaan) yang berbeda, nilai count mentah tidak dapat langsung dibandingkan antar sampel tanpa normalisasi terlebih dahulu. The Cancer Genome Atlas (TCGA) merupakan konsorsium yang menyediakan data genomik kanker berskala besar, dan subset TCGA-PRAD khusus memuat data kanker prostat. Ketersediaan data publik ini memungkinkan penelitian dapat direplikasi dan dibandingkan antar kelompok riset, sekaligus menyediakan ukuran sampel yang memadai untuk analisis ekspresi diferensial maupun pembelajaran mesin. Dalam penelitian ini, matriks ekspresi tersebut menjadi masukan utama bagi tahap seleksi gen dan pemodelan klasifikasi.



Gambar 2.2. Alur Perolehan Data Ekspresi Gen RNA-Seq

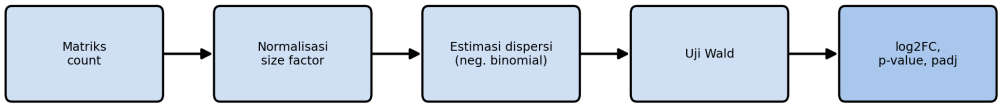
RNA-Sequencing (RNA-Seq) merupakan teknologi sekuensing generasi baru (next-generation sequencing) yang mengukur tingkat ekspresi gen dengan cara menghitung jumlah fragmen transkrip (reads) yang berasal dari tiap gen. Proses ini diawali dengan ekstraksi RNA dari jaringan, konversi menjadi complementary DNA (cDNA), fragmentasi, dan sekuensing masif secara paralel, yang kemudian dipetakan ke genom referensi untuk memperoleh matriks hitungan (count matrix). Dibandingkan teknologi microarray terdahulu, RNA-Seq memiliki rentang dinamis yang lebih lebar, mampu mendeteksi transkrip dengan ekspresi

sangat rendah maupun sangat tinggi, serta tidak bergantung pada probe yang telah ditentukan sebelumnya, sehingga dapat menemukan transkrip baru [6, 12].

The Cancer Genome Atlas (TCGA) merupakan proyek kolaboratif berskala besar yang menyediakan data multi-omik, termasuk RNA-Seq, untuk berbagai jenis kanker beserta metadata klinisnya. Dataset khusus kanker prostat, yaitu TCGA-PRAD (Prostate Adenocarcinoma), menyediakan profil ekspresi puluhan ribu gen untuk ratusan sampel jaringan kanker dan normal. Karena bersifat publik, terstandardisasi, dan berukuran besar, TCGA-PRAD banyak digunakan sebagai tolok ukur dalam penelitian bioinformatika, sehingga hasil penelitian yang menggunakannya dapat dibandingkan secara adil dengan studi lain [6, 12]. Nilai ekspresi pada dataset ini disediakan dalam format ternormalisasi log2(count+1); untuk keperluan analisis ekspresi diferensial dengan DESeq2 yang mengasumsikan data hitungan bulat, nilai tersebut perlu ditransformasi balik menjadi raw counts terlebih dahulu [7].

2.4 Analisis Ekspresi Diferensial dengan DESeq2

DESeq2 adalah metode standar untuk analisis ekspresi diferensial pada data RNA-Seq. DESeq2 membandingkan ekspresi gen antara dua kondisi dan menghasilkan log2 fold change (log2FC), p-value, serta adjusted p-value (padj). Studi pembanding menunjukkan bahwa metode berbasis negative binomial seperti DESeq2 memberikan kontrol false discovery rate dan sensitivitas yang baik [7]. Implementasi yang digunakan pada penelitian ini adalah pydeseq2, yaitu versi Python dari DESeq2. PyDESeq2 menghasilkan model likelihood yang lebih tinggi, memberikan peningkatan kecepatan pada dataset besar seperti TCGA, dan tersedia sebagai perangkat lunak sumber terbuka, sehingga mudah diintegrasikan dengan perkakas data science berbasis Python [8]. Tahapan analisis DESeq2 dirangkum pada Gambar 2.3.

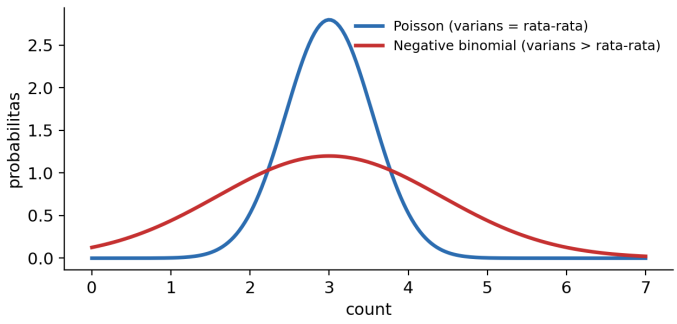


Gambar 2.3. Tahapan Analisis Ekspresi Diferensial DESeq2

2.4.1 Model Distribusi Negative Binomial

Data count RNA-Seq bersifat overdispersi, yaitu varians lebih besar daripada rata-rata, sehingga tidak sesuai dimodelkan dengan distribusi Poisson. DESeq2 memodelkan count menggunakan distribusi negative binomial dengan estimasi dispersi per-gen, lalu menguji signifikansi perbedaan ekspresi menggunakan uji Wald [7]. Perbedaan antara distribusi Poisson dan negative binomial diilustrasikan pada Gambar 2.4.

Distribusi Poisson mengasumsikan varians sama dengan rata-rata, sedangkan pada data biologis nyata terdapat variabilitas biologis antar-sampel yang membuat varians jauh lebih besar. Distribusi negative binomial menambahkan parameter dispersi yang menangkap variabilitas ekstra ini, sehingga uji signifikansi tidak menghasilkan terlalu banyak positif palsu. Karena jumlah sampel per gen umumnya terbatas, estimasi dispersi per-gen cenderung tidak stabil; DESeq2 mengatasinya dengan menyusutkan (shrinkage) estimasi dispersi tiap gen ke arah tren dispersi bersama seluruh gen, sehingga estimasinya lebih andal. Setelah dispersi diestimasi, DESeq2 mencocokkan model linier tergeneralisasi untuk memperoleh log2 fold change tiap gen, lalu menguji apakah perubahan tersebut berbeda signifikan dari nol menggunakan uji Wald. Rangkaian langkah ini menjadikan DESeq2 lebih tepat untuk data count dibandingkan uji statistik klasik yang mengasumsikan data berdistribusi normal.



Gambar 2.4. Perbandingan Distribusi Poisson dan Negative Binomial

Distribusi negative binomial dipilih karena memiliki dua parameter, yaitu rata-rata dan parameter dispersi, sehingga dapat memodelkan variabilitas data hitungan yang melebihi asumsi distribusi Poisson (yang variansnya selalu sama dengan rata-ratanya). Pada data RNA-Seq, variabilitas biologis antar-sampel menyebabkan varians ekspresi suatu gen jauh lebih besar daripada rata-ratanya,

terutama pada gen dengan ekspresi tinggi. DESeq2 mengestimasi parameter dispersi untuk setiap gen, kemudian menstabilkan estimasi tersebut dengan meminjam informasi dari seluruh gen melalui pendekatan penyusutan (shrinkage) menuju sebuah tren dispersi bersama. Mekanisme ini penting ketika jumlah sampel terbatas, karena estimasi dispersi per-gen yang dihitung sendiri-sendiri cenderung tidak stabil dan berujung pada hasil uji yang kurang andal [7].
Setelah parameter model diestimasi, DESeq2 menguji signifikansi perbedaan ekspresi antar-kondisi menggunakan uji Wald terhadap koefisien log2 fold change. Setiap gen memperoleh nilai p yang menyatakan peluang perbedaan sebesar itu muncul secara kebetulan apabila gen sebenarnya tidak berbeda. Karena puluhan ribu gen diuji secara bersamaan, nilai p mentah akan menghasilkan banyak temuan positif palsu, sehingga diperlukan koreksi pengujian ganda. DESeq2 menerapkan prosedur Benjamini-Hochberg untuk mengendalikan false discovery rate (FDR) dan menghasilkan adjusted p-value (padj). Dengan demikian, gen yang lolos ambang padj yang kecil dapat diyakini benar-benar berbeda ekspresinya, bukan sekadar artefak dari banyaknya pengujian yang dilakukan [7, 8].

Nilai log2 fold change yang menjadi dasar penentuan arah dan besar perubahan ekspresi dihitung sebagai logaritma basis dua dari rasio ekspresi rata-rata kondisi kanker terhadap normal, sebagaimana Rumus 2.1.

$$\text{log2FC} = \log_2\left(\frac{\bar{x}_{\text{kanker}}}{\bar{x}_{\text{normal}}}\right) \tag{2.1}$$

2.4.2 Normalisasi dan Koreksi Multiple Testing

DESeq2 melakukan normalisasi size factor untuk mengoreksi perbedaan kedalaman sekuensing antar sampel. Karena ribuan gen diuji secara serentak, DESeq2 menerapkan koreksi Benjamini-Hochberg untuk mengendalikan false discovery rate (FDR) melalui nilai padj [7, 8].

2.4.3 Justifikasi Pemilihan DESeq2 sebagai Metode Seleksi Fitur

Pemilihan DESeq2 sebagai metode seleksi fitur mengikuti panduan literatur terdahulu yang menunjukkan bahwa metode berbasis negative binomial lebih tepat untuk data hitungan RNA-Seq dibandingkan uji statistik umum [7, 8]. Ambang seleksi yang digunakan (padj $<$ 0,05, baseMean $\geq$ 10, dan log2FC $>$ 0,35)

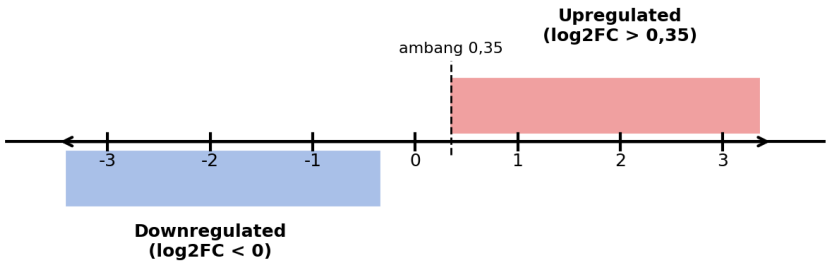
juga merupakan kriteria yang lazim dipakai pada analisis ekspresi diferensial. Perbandingan DESeq2 terhadap beberapa alternatif metode seleksi fitur disajikan pada Tabel 2.3.

Tabel 2.3. Justifikasi Pemilihan DESeq2 sebagai Metode Seleksi Fitur

Metode	**Karakteristik**	**Alasan tidak dipilih / dipilih**
Uji-t / fold change saja	Sederhana, cepat	Tidak memodelkan overdispersi count; rentan positif palsu
edgeR / limma-voom	Berbasis count, populer	Alternatif valid, namun DESeq2 lebih stabil untuk sampel kecil
Seleksi berbasis korelasi	Mudah diterapkan	Tidak memberi signifikansi statistik yang terkontrol FDR
DESeq2 (dipilih)	**Negative binomial, normalisasi size factor, koreksi FDR**	**Standar untuk RNA-Seq, kontrol FDR baik, tersedia di Python (pydeseq2)**

2.5 Gen Upregulated dan Biomarker Kanker Prostat

Gen upregulated adalah gen yang ekspresinya lebih tinggi pada jaringan kanker dibandingkan jaringan normal (log2FC $>$ 0). Beberapa gen yang ditemukan pada penelitian ini telah dikenal sebagai penanda kanker prostat dalam literatur. PCA3 merupakan lncRNA spesifik prostat yang banyak digunakan sebagai biomarker urin; sel kanker prostat diketahui memiliki kadar mRNA PCA3 yang sangat tinggi sehingga menjadikannya salah satu biomarker untuk diagnosis dan prognosis kanker prostat [21]. AMACR telah divalidasi sebagai penanda jaringan kanker prostat [22], dan PCAT14 termasuk lncRNA terkait kanker prostat [23]. Validitas biologis ini memperkuat keandalan tanda tangan gen yang diperoleh. Konsep arah ekspresi berdasarkan nilai log2FC ditunjukkan pada Gambar 2.5.



Gambar 2.5. Interpretasi Nilai Log2 Fold Change

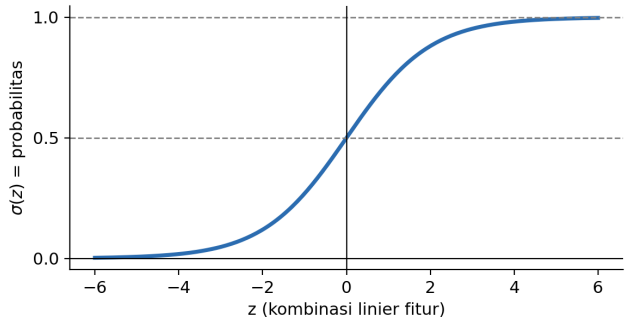
2.6 Klasifikasi Machine Learning

2.6.1 Regresi Logistik

Regresi Logistik adalah model klasifikasi linier yang memodelkan probabilitas suatu sampel termasuk kelas tertentu melalui fungsi sigmoid. Model ini sederhana, mudah diinterpretasikan, dan banyak digunakan sebagai baseline pada klasifikasi berbasis ekspresi gen [24]. Model menghitung skor linier berbobot dari fitur ekspresi gen (Rumus 2.2), lalu memetakannya menjadi probabilitas melalui fungsi sigmoid (Rumus 2.3). Nilai probabilitas di atas 0,5 dikelaskan sebagai kanker. Fungsi sigmoid yang memetakan keluaran linier menjadi probabilitas ditunjukkan pada Gambar 2.6.

$$z = b + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{2.2}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$



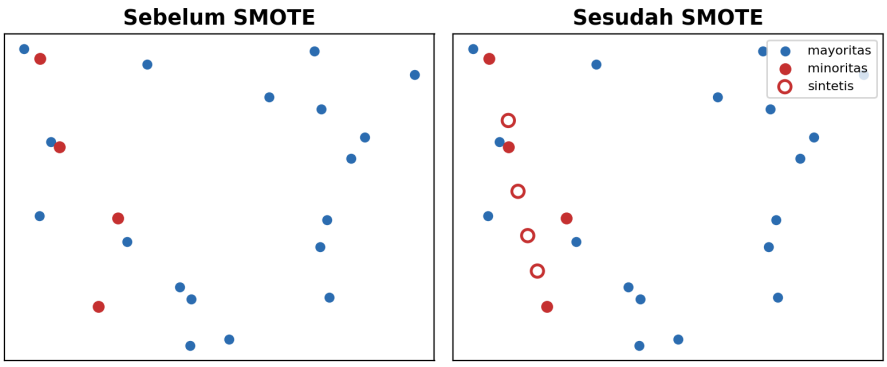
Gambar 2.6. Fungsi Sigmoid pada Regresi Logistik

Secara matematis, Regresi Logistik pertama-tama menghitung kombinasi linier berbobot dari fitur masukan (Rumus 2.2), yang menghasilkan skor z. Skor ini kemudian dipetakan ke rentang probabilitas antara 0 dan 1 melalui fungsi sigmoid (Rumus 2.3). Nilai bobot (w_i) dan bias (b) dipelajari dari data pelatihan dengan cara memaksimalkan fungsi likelihood, yang setara dengan meminimalkan fungsi kerugian log-loss (binary cross-entropy). Proses optimasi dilakukan secara iteratif menggunakan metode berbasis gradien hingga konvergen. Salah satu keunggulan utama Regresi Logistik adalah interpretabilitasnya: besar dan arah tiap bobot menunjukkan seberapa kuat dan ke arah mana suatu gen memengaruhi keputusan klasifikasi, sehingga kontribusi tiap fitur dapat ditelusuri secara langsung [24].
Untuk mencegah pemasangan berlebih (overfitting), Regresi Logistik dilengkapi mekanisme regularisasi yang menambahkan penalti terhadap besarnya bobot pada fungsi kerugian. Terdapat dua jenis penalti yang umum digunakan. Penalti L2 (ridge) menyusutkan seluruh bobot secara halus mendekati nol tanpa menghilangkannya, sehingga cocok ketika seluruh fitur dianggap informatif. Penalti L1 (lasso) dapat memaksa sebagian bobot menjadi tepat nol, sehingga sekaligus melakukan seleksi fitur. Kekuatan regularisasi diatur oleh parameter C yang merupakan kebalikan dari besarnya penalti: nilai C kecil memberikan regularisasi kuat (model lebih sederhana), sedangkan nilai C besar memberikan regularisasi lemah (model lebih fleksibel). Pemilihan kombinasi C dan jenis penalti yang tepat menjadi krusial untuk memperoleh keseimbangan antara kesederhanaan dan kemampuan model [24, 27].

2.6.2 SMOTE untuk Ketidakseimbangan Kelas

Karena jumlah sampel normal jauh lebih sedikit daripada sampel kanker, data bersifat tidak seimbang. Synthetic Minority Over-sampling Technique

(SMOTE) menyeimbangkan data dengan membuat sampel sintetis kelas minoritas. Penanganan ketidakseimbangan yang tepat penting agar model tidak bias terhadap kelas mayoritas [25]. Prinsip kerja SMOTE diilustrasikan pada Gambar 2.7.



Gambar 2.7. Ilustrasi Pembuatan Sampel Sintetis SMOTE

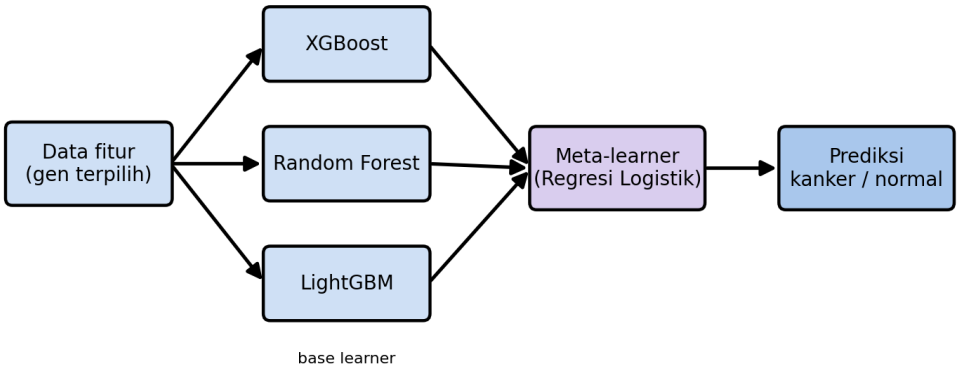
Ketidakseimbangan kelas merupakan permasalahan umum pada data biomedis, termasuk pada penelitian ini di mana jumlah sampel normal (44) jauh lebih sedikit dibandingkan sampel kanker (414). Apabila tidak ditangani, model klasifikasi cenderung bias terhadap kelas mayoritas dan memperoleh akurasi tinggi secara semu dengan mengabaikan kelas minoritas. SMOTE (Synthetic Minority Over-sampling Technique) mengatasi hal ini bukan dengan menggandakan sampel minoritas yang sudah ada, melainkan dengan membangkitkan sampel sintetis baru. Untuk setiap sampel minoritas, SMOTE memilih salah satu tetangga terdekatnya dari kelas yang sama, lalu menempatkan sampel baru pada suatu titik acak di sepanjang garis yang menghubungkan keduanya. Dengan cara interpolasi ini, ruang fitur kelas minoritas menjadi lebih padat dan bervariasi tanpa sekadar menduplikasi data, sehingga model dapat mempelajari batas keputusan kelas minoritas dengan lebih baik.

Hal yang sangat penting dalam penerapan SMOTE adalah penempatannya di dalam alur validasi silang. Apabila SMOTE diterapkan pada seluruh data sebelum pembagian lipatan, sampel sintetis yang dibangkitkan dapat memuat informasi dari sampel yang kelak menjadi data uji, sehingga menimbulkan kebocoran data dan estimasi performa yang terlalu optimis. Oleh karena itu, pada penelitian ini SMOTE ditempatkan di dalam pipeline yang hanya dilatih pada bagian data latih setiap lipatan, sehingga data uji tetap murni dan tidak pernah memengaruhi proses penyeimbangan kelas [25, 17].

2.6.3 Stacking Ensemble

Stacking ensemble menggabungkan beberapa model dasar (base learner), pada penelitian ini XGBoost, Random Forest, dan LightGBM, dengan sebuah meta-learner (Regresi Logistik) yang mempelajari kombinasi terbaik dari prediksi base learner. Ensemble berbasis seleksi fitur dilaporkan efektif untuk klasifikasi kanker berbasis ekspresi gen [20]. Arsitektur stacking ensemble ditunjukkan pada Gambar 2.8.

Ketiga base learner tersebut memiliki cara kerja yang berbeda. Random Forest membangun banyak pohon keputusan secara paralel dari sampel data dan fitur acak (bagging), lalu menggabungkan hasilnya melalui pemungutan suara sehingga mengurangi variansi. XGBoost membangun pohon secara bertahap (boosting), di mana tiap pohon baru memperbaiki kesalahan pohon sebelumnya menggunakan optimasi gradien dengan regularisasi. LightGBM juga menerapkan gradient boosting, tetapi tumbuh secara leaf-wise dan memakai teknik histogram sehingga jauh lebih cepat dan efisien memori pada data berdimensi tinggi. Perbedaan mekanisme inilah yang membuat gabungan ketiganya melalui stacking berpotensi saling melengkapi.



Gambar 2.8. Arsitektur Stacking Ensemble Tiga Base Learner

2.6.4 Dasar Pembandingan dan Justifikasi Pemilihan Algoritma

Kedua pendekatan klasifikasi pada penelitian ini dibandingkan atas dasar yang jelas, yaitu untuk menguji apakah model sederhana yang transparan dapat menyamai atau bahkan mengungguli model ensemble yang kompleks pada permasalahan berdimensi rendah (tujuh gen). Regresi Logistik mewakili model

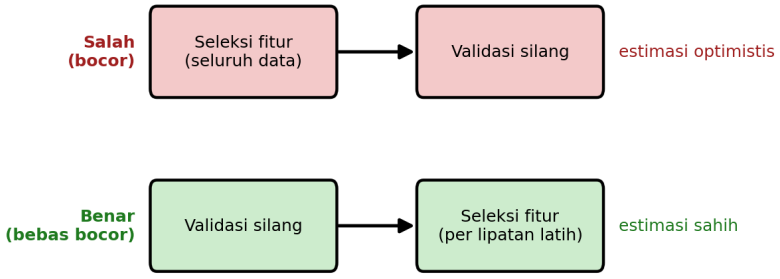
linier yang sederhana dan interpretable, sedangkan stacking ensemble mewakili model non-linier yang kompleks namun kurang transparan. Pembandingan dilakukan pada skema evaluasi yang sama (validasi silang per-pasien bebas kebocoran) dan metrik yang sama (ROC AUC, recall, F1-score), sehingga perbedaan performa murni berasal dari algoritma. Dasar pembandingan ini sejalan dengan prinsip parsimoni, yaitu memilih model paling sederhana yang memberikan performa memadai. Justifikasi pemilihan Regresi Logistik sebagai model utama terhadap alternatif lain dirangkum pada Tabel 2.4.

Tabel 2.4. Justifikasi Pemilihan Algoritma Klasifikasi

Algoritma	**Karakteristik**	**Alasan tidak dipilih / dipilih**
Deep learning	Akurasi tinggi pada data besar	Butuh data besar, rawan overfitting & kurang transparan untuk 7 fitur
Stacking ensemble	Kuat menangkap pola non-linier	Kompleks, kurang interpretable; dipakai sebagai pembanding
SVM / KNN	Baseline umum	Kurang interpretable dibanding model linier berbobot
Regresi Logistik (dipilih)	**Linier, transparan, bobot dapat diinterpretasi**	**Sesuai untuk 7 fitur, sederhana, performa unggul, parsimoni**

2.7 Validasi Silang dan Kebocoran Data

Validasi silang (cross-validation) membagi data menjadi beberapa lipatan untuk memperoleh estimasi performa yang lebih stabil. Namun, jika praproses atau seleksi fitur dilakukan menggunakan seluruh data sebelum pembagian, terjadi kebocoran data yang menyebabkan estimasi performa terlalu optimis [17, 18]. Sebagai bukti, sebuah studi pada sepuluh dataset publik menunjukkan bahwa melakukan seleksi fitur sebelum validasi silang menghasilkan estimasi performa yang bias dibandingkan ketika seleksi dilakukan dengan benar di dalam validasi silang [17]. Pada data genomika, kebocoran ini merupakan kesalahan umum yang perlu dihindari [19]. Oleh karena itu, seleksi gen dan penyeimbangan kelas harus dilakukan di dalam setiap lipatan pelatihan, dan pembagian dilakukan per-pasien. Perbedaan antara seleksi fitur yang menimbulkan kebocoran dan yang benar diilustrasikan pada Gambar 2.9.



Gambar 2.9. Perbedaan Seleksi Fitur yang Bocor dan yang Sahih

Validasi silang k-fold membagi data menjadi k bagian yang kemudian digunakan secara bergantian sebagai data uji, sementara sisanya menjadi data latih, sehingga setiap sampel pernah menjadi data uji tepat satu kali. Rata-rata dari k hasil pengukuran memberikan estimasi performa yang lebih stabil dibandingkan pembagian tunggal. Pada data biomedis, terdapat dua jenis kebocoran yang perlu diwaspadai. Pertama, kebocoran pada seleksi fitur, yang terjadi apabila pemilihan gen dilakukan menggunakan seluruh data termasuk data uji, sehingga fitur terpilih telah ”melihat” jawaban. Kedua, kebocoran tingkat kelompok, yang terjadi apabila sampel-sampel yang berasal dari pasien yang sama tersebar pada data latih maupun data uji, sehingga model dapat mengenali pasien alih-alih mempelajari pola penyakit.

Untuk mengatasi kedua sumber kebocoran tersebut, penelitian ini menggunakan skema StratifiedGroupKFold. Komponen ”Stratified” menjaga agar proporsi kelas kanker dan normal tetap seimbang pada setiap lipatan, yang penting mengingat sedikitnya jumlah sampel normal. Komponen ”Group” memastikan seluruh sampel milik satu pasien selalu berada pada lipatan yang sama, sehingga tidak ada pasien yang muncul di data latih sekaligus data uji. Dengan menggabungkan kedua sifat tersebut serta menempatkan seleksi gen dan SMOTE di dalam setiap lipatan pelatihan, estimasi performa yang dihasilkan menjadi jujur dan mencerminkan kemampuan generalisasi model pada pasien yang benar-benar baru [17, 19].

2.8 Metrik Evaluasi

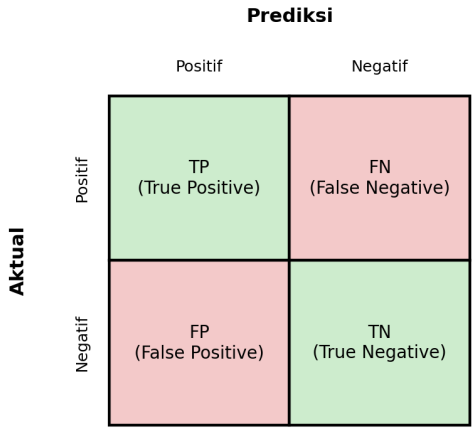
Karena data sangat tidak seimbang, akurasi saja tidak cukup karena model yang selalu menebak kelas mayoritas pun dapat memperoleh akurasi tinggi. ROC

AUC serta metrik berbasis kelas minoritas seperti recall dan F1-score lebih informatif untuk dataset tidak seimbang [26]. Penelitian ini melaporkan ROC AUC, F1-score, recall, precision, dan accuracy secara bersamaan. Precision, recall, dan F1-score dihitung dari komponen confusion matrix (true positive TP, false positive FP, dan false negative FN) sebagaimana Rumus 2.4–2.6. Dasar perhitungan metrik-metrik tersebut, yaitu confusion matrix, ditunjukkan pada Gambar 2.10.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.6}$$



Gambar 2.10. Confusion Matrix sebagai Dasar Perhitungan Metrik

Confusion matrix merupakan tabel dua dimensi yang membandingkan label prediksi terhadap label sebenarnya, terdiri atas true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). Dari keempat komponen ini diturunkan berbagai metrik. Accuracy mengukur proporsi prediksi yang benar secara keseluruhan, tetapi menjadi menyesatkan pada data tidak seimbang karena model yang selalu menebak kelas mayoritas pun dapat memperoleh accuracy tinggi. Precision (Rumus 2.4) mengukur ketepatan prediksi positif, sedangkan

recall (Rumus 2.5) mengukur kemampuan menangkap seluruh kasus positif yang sebenarnya. F1-score (Rumus 2.6) merupakan rata-rata harmonik precision dan recall yang menyeimbangkan keduanya.

Selain metrik berbasis ambang tunggal tersebut, ROC AUC (Area Under the Receiver Operating Characteristic Curve) mengukur kemampuan model membedakan kelas pada seluruh kemungkinan ambang keputusan. Kurva ROC menggambarkan hubungan antara true positive rate dan false positive rate ketika ambang divariasikan, dan luas di bawah kurva tersebut merangkum performa dalam satu angka antara 0,5 (setara tebakan acak) hingga 1,0 (pemisahan sempurna). Karena tidak bergantung pada ambang tertentu dan relatif tahan terhadap ketidakseimbangan kelas, ROC AUC menjadi metrik utama pada penelitian ini, dilengkapi recall dan F1-score yang lebih sensitif terhadap kelas minoritas. Kombinasi metrik ini memberikan gambaran performa yang lebih menyeluruh dan jujur dibandingkan hanya melaporkan accuracy [26].

2.9 Interpretabilitas Model dan Relevansi Klinis

Dalam penerapan machine learning pada bidang biomedis, interpretabilitas model menjadi pertimbangan yang sama pentingnya dengan akurasi. Model yang bersifat kotak hitam (black box), meskipun akurat, sulit diterima dalam praktik klinis karena tenaga medis perlu memahami dasar dari setiap keputusan sebelum menggunakannya untuk pasien. Regresi Logistik memiliki keunggulan dalam hal ini karena setiap fitur diberi bobot yang dapat dibaca secara langsung: tanda bobot menunjukkan arah pengaruh gen terhadap probabilitas kanker, sedangkan besarnya menunjukkan kekuatan pengaruh tersebut. Dengan demikian, kontribusi tiap gen terhadap keputusan klasifikasi dapat ditelusuri dan diaudit, yang sangat berharga ketika model hendak dijelaskan kepada klinisi maupun ketika hasil model perlu dipertanggungjawabkan.

Relevansi klinis juga ditentukan oleh ukuran dan kepraktisan panel penanda. Panel yang terdiri atas sejumlah kecil gen jauh lebih mudah diterjemahkan menjadi uji diagnostik yang terjangkau dibandingkan model yang bergantung pada ribuan gen. Uji berbasis sedikit gen dapat dikembangkan menggunakan teknologi yang lebih sederhana dan murah, sehingga berpotensi diterapkan pada fasilitas kesehatan dengan sumber daya terbatas. Oleh karena itu, kombinasi antara panel gen yang ringkas dan model yang interpretable tidak hanya bernilai secara ilmiah, tetapi juga membuka peluang penerapan yang lebih luas, sejalan dengan tujuan akhir

penelitian biomarker, yaitu menghadirkan alat bantu diagnosis yang objektif, terjangkau, dan dapat dipercaya.