BAB 2
LANDASAN TEORI

2.1 Tinjauan Teori

Bagian 2.1.1 menjelaskan secara singkat tentang kanker usus besar, informasi penentuan stadium yang berasal dari sistem AJCC. Bagian 2.1.2 menjelaskan fungsi ANOVA untuk melakukan filter dalam memilih gen dengan variansi tinggi. Bagian 2.1.3 memaparkan kerangka seleksi fitur berbasis approximate conditional entropy, mulai dari definisi sistem informasi keputusan, relasi *fuzzy* berbasis *Laplacian kernel, approximate accuracy, approximate conditional entropy* (H_{ace}), hingga konsep reduksi atribut. Bagian ini juga menjelaskan cara mengevaluasi pentingnya setiap atribut menggunakan indikator *Importance of Internal Attribute* (IIA) dan *Importance of External Attribute* (IEA), serta alur iteratif seleksi fitur menggunakan FSACE. Terakhir, bagian 2.1.4 menguraikan model klasifikasi XGBoost yang digunakan untuk menilai kinerja subset fitur terpilih, termasuk fungsi objektif dan mekanisme pencegahan *overfitting*.

2.1.1 Kanker Usus Besar dan Staging

Kanker usus besar (*colon cancer*) merupakan salah satu jenis kanker yang paling umum dan mematikan di dunia. Kanker ini berasal dari sel epitel pada dinding usus besar, termasuk kolon dan rektum. Secara klinis, kanker usus besar dapat menimbulkan gejala seperti perubahan kebiasaan buang air besar, perdarahan rektal, nyeri perut, penurunan berat badan, dan anemia.

Untuk menentukan stadium kanker dan rencana terapi, sistem *staging* AJCC (*American Joint Committee on Cancer*) edisi ke-8 menggunakan kombinasi tiga komponen utama: Tumor (T), Node (N), dan Metastasis (M), yang selanjutnya digabungkan menjadi stage 0 sampai IV. Komponen TNM dijelaskan sebagai berikut:

1. Tumor (T): Ukuran dan penetrasi tumor pada dinding usus.
   (a) Tis: *Carcinoma in situ*, sel kanker hanya berada pada lapisan mukosa.
   (b) T1: Tumor menembus submukosa.
   (c) T2: Tumor menembus *muscularis propria*.
   (d) T3: Tumor menembus *muscularis propria* ke jaringan sekitar.
   (e) T4a: Tumor menembus peritoneum visceral.
   (f) T4b: Tumor menempel atau menyusup ke organ/struktur lain.
2. Node (N): Keterlibatan kelenjar getah bening regional.
   (a) N0: Tidak ada kelenjar getah bening positif.
   (b) N1: 1–3 kelenjar getah bening positif.
   (c) N2: 4 atau lebih kelenjar getah bening positif.

3. Metastasis (M): Penyebaran ke organ jauh.

   (a) M0: Tidak ada metastasis jauh.

   (b) M1a: Metastasis tunggal ke organ lain.

   (c) M1b: Metastasis ke lebih dari satu organ.

   (d) M1c: Metastasis termasuk peritoneum.

Berdasarkan kombinasi TNM tersebut, klasifikasi stadium kanker usus besar menurut AJCC edisi ke-8 disajikan pada Tabel 2.1.

Tabel 2.1. Pengelompokan Stadium Kanker Usus Besar menurut AJCC Edisi ke-8

Stage	Tumor (T)	Node (N), Metastasis (M)
0	Tis	N0 M0
I	T1, T2	N0 M0
IIA	T3	N0 M0
IIB	T4a	N0 M0
IIC	T4b	N0 M0
IIIA	T1-2	N1 M0
IIIB	T3-4a	N1 M0
IIIC	Any T	N2 M0
IVA	Any T	Any N M1a
IVB	Any T	Any N M1b
IVC	Any T	Any N M1c

Sumber: *American Joint Committee on Cancer* (AJCC) [2]

Klasifikasi ini penting untuk menentukan prognosis dan strategi pengobatan, di mana stadium lebih tinggi (misal IV) menunjukkan penyebaran lebih luas dan memerlukan intervensi yang lebih agresif dibanding stadium awal (misal I atau II).

2.1.2 ANOVA (Analysis of Variance)

Analysis of Variance (ANOVA) merupakan salah satu teknik statistik klasik yang digunakan untuk menguji apakah terdapat perbedaan berarti pada rata-rata sebuah variabel kontinu di antara dua atau lebih kelompok kategori. Secara umum, ANOVA menghitung nilai F dengan membandingkan varians antar kelompok terhadap varians dalam kelompok. Perbandingan ini menunjukkan sejauh mana variasi pada suatu fitur (misalnya ekspresi gen) berkaitan dengan perbedaan kelas target — dalam penelitian klasifikasi kanker ekspresi gen, kelas target ini dapat berupa stadium kanker atau kondisi biologis tertentu.

Secara matematis, statistik F dalam ANOVA dinyatakan sebagai rasio antara varians antar kelas dan varians dalam kelas. Nilai F yang lebih tinggi menunjukkan bahwa perbedaan rata-rata ekspresi gen antar kelas lebih dominan dibanding variasi dalam kelas itu sendiri.

Dalam konteks seleksi fitur untuk dataset ekspresi gen kanker, ANOVA berperan sebagai metode *filter*. Metode filter bekerja dengan mengevaluasi setiap gen secara independen terhadap kelas target tanpa keterikatan langsung dengan model klasifikasi tertentu. Penggunaan ANOVA memungkinkan identifikasi gen-gen yang memiliki perbedaan ekspresi signifikan antar kelas tumor dibanding kelas lainnya, sehingga gen-gen tersebut dianggap sebagai kandidat fitur yang informatif dan relevan untuk pengklasifikasian lebih lanjut. Strategi ini efektif dalam mereduksi ruang fitur yang sangat besar sebelum menerapkan teknik seleksi fitur multivariat atau pembelajaran mesin yang lebih kompleks [3].

ANOVA juga dipahami sebagai salah satu tes statistik univariat yang sering dipakai dalam studi gen ekspresi untuk memilih fitur-fitur awal yang potensial. Seperti dijelaskan oleh Abduallah (2023), ANOVA digunakan sebagai salah satu algoritma seleksi fitur pada profil ekspresi gen untuk meningkatkan akurasi klasifikasi kanker, yang menyoroti peran penting ANOVA dalam tahap awal pemilihan fitur untuk dataset gen ekspresi yang berdimensi tinggi [3].

Singkatnya, ANOVA memberi ukuran statistik yang menyatakan apakah ekspresi suatu gen berbeda secara signifikan di antara kelas target yang berbeda, sehingga membantu menyaring gen-gen yang lebih mungkin berkaitan dengan perbedaan biologis antar kelas sebelum langkah seleksi lanjutan dilakukan. Pendekatan ini membantu mengurangi jumlah fitur yang dipertimbangkan, menurunkan kompleksitas komputasi, serta meningkatkan fokus pada gen-gen yang paling berpotensi informatif dalam konteks klasifikasi kanker berdasarkan ekspresi gen.

2.1.3 Feature Selection Using Approximate Conditional Entropy

Dalam sistem informasi keputusan $IS = (U, C \cup D)$, dengan U himpunan objek (pasien), C himpunan atribut kondisi (fitur ekspresi gen), dan D atribut keputusan (kelas), *feature selection* dilakukan untuk mengurangi dimensi data sekaligus mempertahankan informasi penting terkait keputusan. Untuk menangani data kontinu dan ketidakpastian, digunakan pendekatan *fuzzy set* dengan relasi kemiripan *Laplacian kernel* [1]:

$$R_c(x_i, x_j) = \exp\left(-\frac{\|c(x_i) - c(x_j)\|}{\sigma_c}\right) \tag{2.1}$$

Relasi fuzzy ini membentuk *information granule* untuk setiap objek, yang menjadi dasar perhitungan *approximate accuracy* dan *approximate conditional entropy* (H_{ace}). Approximate accuracy mengukur ketidakpastian atau ketidaktepatan granule informasi yang dibentuk dari subset atribut $B \subseteq C$. Untuk $X \subseteq U, X \neq \emptyset$, approximate accuracy didefinisikan sebagai:

$$a_B(X) = \frac{|P(X)|}{|\overline{P}(X)|}, \quad 0 \leq a_B(X) \leq 1 \tag{2.2}$$

di mana $P(X)$ adalah *positive region* dari X dan $|\cdot|$ menunjukkan kardinalitas himpunan.

Berdasarkan approximate accuracy, *approximate conditional entropy* H_{ace} mengukur ketidakpastian informasi dari atribut keputusan D relatif terhadap subset atribut B. Jika U berasal

dari D = $\{X_1, X_2, \ldots, X_k\}$ adalah partisi objek menurut keputusan, maka H_{ace} didefinisikan sebagai:

$$H_{ace}(D|B) = -\sum_{j=1}^{k}\sum_{i=1}^{|U|}\log(2 - a_B(X_j))\frac{|[x_i]_{R_B} \cap X_j|}{|[x_i]_{R_B}|} \tag{2.3}$$

H_{ace} menggabungkan efek ketidakpastian *granule* informasi dan ketidaktepatan *boundary region*, sehingga memberikan ukuran yang lebih lengkap terhadap informasi atribut.

Subset atribut $B \subseteq C$ dikatakan sebagai *reduction* dari C relatif terhadap D jika memenuhi dua kondisi [1]: pertama, $H_{ace}(D|B) = H_{ace}(D|C)$, artinya subset terpilih memiliki jumlah informasi sama dengan seluruh atribut; kedua, $H_{ace}(D|B - \{b\}) > H_{ace}(D|C)$ untuk semua $b \in B$, memastikan tidak ada redundansi dalam subset.

Untuk mengevaluasi kontribusi masing-masing atribut dalam subset penuh C, digunakan *Importance of Internal Attribute* (IIA):

$$IIA(c, C, D) = H_{ace}(D|C - \{c\}) - H_{ace}(D|C) \tag{2.4}$$

Jika $IIA(c, C, D) > 0$, atribut c dianggap sebagai *core attribute*.

Untuk mengevaluasi atribut kandidat $d \in C - B$ yang belum masuk ke subset B, digunakan *Importance of External Attribute* (IEA):

$$IEA(d, B, C, D) = H_{ace}(D|B) - H_{ace}(D|B \cup \{d\}) \tag{2.5}$$

Atribut dengan nilai IEA yang lebih tinggi memiliki kontribusi lebih signifikan terhadap penurunan ketidakpastian informasi, sehingga menjadi prioritas untuk dimasukkan ke subset fitur.

Proses seleksi fitur menggunakan FSACE dilakukan secara iteratif dengan pendekatan *forward selection*, menambahkan fitur berdasarkan nilai IEA tertinggi hingga tidak ada penurunan H_{ace} yang signifikan atau jumlah fitur mencapai batas tertentu [1].

2.1.4 Model Klasifikasi XGBoost

XGBoost (*eXtreme Gradient Boosting*) merupakan algoritma pembelajaran mesin berbasis *gradient boosting* yang membangun model prediksi secara aditif menggunakan kumpulan pohon keputusan (*decision trees*). Algoritma ini dirancang untuk mengoptimalkan kinerja prediksi melalui pemodelan kesalahan residual secara iteratif, sekaligus mengendalikan kompleksitas model menggunakan mekanisme regularisasi yang eksplisit [4].

Misalkan diberikan dataset pelatihan $\{(x_i, y_i)\}_{i=1}^{n}$, dengan $x_i \in \mathbb{R}^m$ menyatakan vektor fitur dan y_i label kelas. Model XGBoost memprediksi keluaran $\hat{y}_i$ sebagai penjumlahan dari K fungsi pohon keputusan:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathscr{F} \tag{2.6}$$

dengan $\mathscr{F}$ adalah ruang fungsi pohon keputusan yang memetakan fitur ke bobot pada simpul daun.

Proses pembelajaran XGBoost bertujuan meminimalkan fungsi objektif berikut:

$$\mathscr{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.7}$$

di mana $\ell(\cdot)$ adalah fungsi kerugian (*loss function*), sedangkan $\Omega(f_k)$ merupakan fungsi regularisasi yang mengontrol kompleksitas model.

Untuk sebuah pohon keputusan f, fungsi regularisasi didefinisikan sebagai:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.8}$$

dengan T jumlah simpul daun pada pohon, w_j bobot pada daun ke-j, γ penalti kompleksitas struktur pohon, dan λ parameter regularisasi bobot. Regularisasi ini berperan penting dalam mencegah *overfitting*, khususnya pada data berdimensi tinggi seperti ekspresi gen.

Pada setiap iterasi, XGBoost membangun pohon baru untuk mempelajari kesalahan residual model sebelumnya menggunakan pendekatan optimisasi orde kedua. Dengan melakukan aproksimasi Taylor hingga orde kedua terhadap fungsi kerugian, optimisasi dilakukan menggunakan gradien pertama dan kedua (Hessian), sehingga proses pembelajaran menjadi lebih stabil dan efisien dibandingkan metode *gradient boosting* konvensional.

Dalam konteks klasifikasi biner, XGBoost umumnya menggunakan fungsi objektif *logistic loss* untuk memodelkan probabilitas kelas. Kombinasi optimisasi berbasis gradien, struktur pohon keputusan, serta regularisasi yang eksplisit menjadikan XGBoost sangat efektif dalam menangani data dengan jumlah fitur besar, korelasi kompleks, dan rasio fitur terhadap sampel yang tinggi.

Berbagai penelitian terkini menunjukkan bahwa XGBoost mampu memodelkan hubungan non-linear antar fitur gen dan label klinis secara efektif, sekaligus menyediakan mekanisme evaluasi kontribusi fitur melalui struktur pohon keputusan yang dibangun [5]. Oleh karena itu, XGBoost digunakan dalam penelitian ini sebagai model klasifikasi untuk mengevaluasi kualitas subset fitur hasil seleksi FSACE.