

## **BAB 2**

### **TINJAUAN PUSTAKA**

#### **2.1 Kanker Lambung (Gastric Cancer)**

Kanker lambung atau *gastric cancer* merupakan keganasan yang berasal dari pertumbuhan sel abnormal pada lapisan mukosa lambung. Penyakit ini termasuk salah satu penyebab utama kematian akibat kanker di dunia. Berdasarkan laporan *Global Cancer Observatory (GLOBOCAN)*, kanker lambung menempati peringkat kelima dalam insidensi kanker global dan peringkat keempat sebagai penyebab kematian akibat kanker dengan lebih dari satu juta kasus baru setiap tahunnya [1].

Tingginya angka mortalitas kanker lambung terutama disebabkan oleh keterlambatan diagnosis, karena gejala awal yang bersifat tidak spesifik, seperti nyeri ulu hati, mual, dan penurunan nafsu makan. Akibatnya, sebagian besar pasien terdiagnosis pada stadium lanjut, sehingga pilihan terapi menjadi terbatas dan prognosis cenderung buruk [7].

Secara biologis, perkembangan kanker lambung mengikuti proses bertahap yang dikenal sebagai *Correa cascade*, yang dimulai dari gastritis kronis, diikuti atrofi mukosa, metaplasia intestinal, displasia, hingga berkembang menjadi adenokarsinoma [8]. Adenokarsinoma merupakan tipe kanker lambung yang paling umum dan mencakup lebih dari 90% dari seluruh kasus kanker lambung.

Faktor risiko utama kanker lambung meliputi infeksi *Helicobacter pylori*, predisposisi genetik, konsumsi makanan tinggi garam, kebiasaan merokok, konsumsi alkohol, serta paparan faktor lingkungan tertentu [9].

##### **2.1.1 Stadium Kanker Lambung**

Penentuan stadium kanker lambung memiliki peranan penting dalam menentukan strategi terapi dan prognosis pasien. Sistem stadium yang paling banyak digunakan adalah sistem TNM (*Tumor, Node, Metastasis*) yang dikembangkan oleh *American Joint Committee on Cancer (AJCC)* [2].

Pada sistem TNM, stadium kanker ditentukan berdasarkan tiga komponen utama, yaitu ukuran dan kedalaman invasi tumor primer (T), keterlibatan kelenjar getah bening regional (N), serta keberadaan metastasis jauh (M). Secara klinis, kanker lambung diklasifikasikan ke dalam empat stadium utama, yaitu stadium I

hingga stadium IV.

Stadium awal (stadium I dan II) umumnya memiliki tingkat kelangsungan hidup yang lebih baik dan memungkinkan dilakukan tindakan kuratif seperti pembedahan. Sebaliknya, stadium lanjut (stadium III dan IV) sering kali memerlukan terapi sistemik seperti kemoterapi atau imunoterapi dengan tingkat keberhasilan yang relatif lebih rendah [10].

### **2.1.2 Keterbatasan Metode Konvensional**

Penentuan stadium kanker lambung secara konvensional umumnya mengandalkan pemeriksaan histopatologi, pencitraan medis, dan evaluasi klinis. Meskipun pendekatan ini telah menjadi standar emas dalam praktik klinis, metode tersebut bersifat invasif dan sangat bergantung pada interpretasi subjektif tenaga medis. Selain itu, pendekatan konvensional belum sepenuhnya mampu menangkap heterogenitas biologis tumor pada tingkat molekuler, yang berperan penting dalam progresi penyakit [3].

Keterbatasan tersebut mendorong pengembangan pendekatan berbasis biomarker molekuler sebagai alternatif atau pelengkap metode konvensional. Salah satu pendekatan yang banyak dikembangkan adalah analisis ekspresi gen yang memungkinkan pengamatan langsung terhadap aktivitas biologis sel kanker. Pendekatan ini diharapkan mampu memberikan klasifikasi stadium kanker yang lebih objektif, presisi, dan bersifat personal.

## **2.2 Analisis Ekspresi Gen**

Analisis ekspresi gen (*gene expression analysis*) bertujuan untuk mengukur tingkat aktivitas gen dalam suatu kondisi biologis tertentu. Data ekspresi gen umumnya diperoleh melalui teknologi *high-throughput* seperti *RNA sequencing* (*RNA-seq*) yang mampu mengukur ribuan gen secara simultan. Teknologi ini menghasilkan data berdimensi tinggi dengan jumlah fitur yang sangat besar pada setiap sampel.

Dalam konteks kanker, perbedaan pola ekspresi gen mencerminkan perubahan biologis yang berkaitan dengan progresi tumor dan agresivitas sel kanker. Profil ekspresi gen juga berhubungan erat dengan karakteristik molekuler yang membedakan setiap stadium penyakit. Studi dari *The Cancer Genome Atlas* (*TCGA*) menunjukkan bahwa pola ekspresi gen memiliki keterkaitan yang kuat

dengan karakteristik klinis dan molekuler kanker lambung [3].

Namun, rasio jumlah fitur yang jauh lebih besar dibandingkan jumlah sampel menimbulkan tantangan dalam analisis data. Kondisi ini dapat menyebabkan fenomena *curse of dimensionality* yang berdampak pada penurunan performa model. Selain itu, risiko terjadinya *overfitting* juga meningkat apabila tidak dilakukan penanganan yang tepat.

### 2.3 Seleksi Fitur pada Data Ekspresi Gen

Seleksi fitur merupakan tahapan penting dalam analisis data berdimensi tinggi yang bertujuan untuk mengurangi jumlah fitur dengan tetap mempertahankan informasi yang paling relevan [4]. Proses ini membantu menghilangkan fitur yang redundan maupun tidak informatif. Dengan demikian, kompleksitas model dapat dikurangi tanpa mengorbankan performa klasifikasi.

Selain meningkatkan performa model, seleksi fitur juga berperan dalam meningkatkan efisiensi komputasi. Pengurangan jumlah fitur memungkinkan proses pelatihan model menjadi lebih cepat dan stabil. Lebih jauh lagi, seleksi fitur meningkatkan interpretabilitas biologis dengan mengidentifikasi gen-gen yang berkontribusi signifikan terhadap fenomena yang diamati.

### 2.4 Seleksi Fitur Menggunakan LASSO

Seleksi fitur merupakan tahapan penting dalam analisis data berdimensi tinggi yang bertujuan untuk mengurangi jumlah fitur dengan tetap mempertahankan informasi yang paling relevan [4]. Pada penelitian ini, metode seleksi fitur yang digunakan adalah *Least Absolute Shrinkage and Selection Operator (LASSO)*. Metode ini dipilih karena kemampuannya melakukan seleksi fitur dan regularisasi secara simultan.

*LASSO* merupakan metode regresi ter-regularisasi yang menggunakan penalti L1 untuk mengecilkan nilai koefisien fitur. Koefisien fitur yang dianggap kurang relevan akan ditekan hingga bernilai nol sehingga fitur tersebut dieliminasi dari model. Dengan mekanisme ini, *LASSO* mampu menghasilkan model yang lebih sederhana dan mudah diinterpretasikan.

Fungsi objektif *LASSO* dirumuskan sebagai berikut:

$$\min_{\beta} (||y - X\beta||_2^2 + \lambda ||\beta||_1) \quad (2.1)$$

Pendekatan ini sangat sesuai untuk data ekspresi gen yang berdimensi tinggi dan memiliki korelasi antar fitur. Selain itu, *LASSO* mampu meningkatkan interpretabilitas biologis dengan mengidentifikasi gen-gen yang paling berkontribusi terhadap klasifikasi stadium kanker lambung [11, 12].

## 2.5 Machine Learning untuk Klasifikasi Stadium Kanker Lambung

*Machine learning* merupakan pendekatan komputasional yang memungkinkan sistem belajar dari data untuk menghasilkan prediksi secara otomatis [5]. Pendekatan ini banyak digunakan dalam bidang bioinformatika untuk menangani data kompleks dan berdimensi tinggi. Dengan memanfaatkan pola yang terdapat dalam data, *machine learning* mampu memberikan prediksi yang akurat dan konsisten.

*Pipeline* yang digunakan dalam penelitian ini meliputi tahapan pra-pemrosesan data, seleksi fitur, pelatihan model, serta evaluasi performa. Setiap tahapan dirancang untuk memastikan kualitas data dan hasil prediksi yang optimal. Pendekatan ini diharapkan mampu menghasilkan model klasifikasi stadium kanker lambung yang andal.

Model yang digunakan meliputi *Logistic Regression*, *Support Vector Machine (SVM)*, *Random Forest*, dan *Multi-Layer Perceptron (MLP)*. Pemilihan beberapa model dilakukan untuk membandingkan performa dan karakteristik masing-masing algoritma. Dengan demikian, model terbaik dapat ditentukan berdasarkan hasil evaluasi yang komprehensif.

## 2.6 Support Vector Machine

*Support Vector Machine (SVM)* merupakan algoritma klasifikasi yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan data dari kelas yang berbeda. *Hyperplane* ini ditentukan dengan memaksimalkan margin antar kelas. Pendekatan ini membuat *SVM* memiliki kemampuan generalisasi yang baik.

Fungsi optimasi *SVM* dirumuskan sebagai berikut:

$$\min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^n \xi_i \quad (2.2)$$

Penggunaan *kernel non-linear* seperti *Radial Basis Function (RBF)* memungkinkan *SVM* menangkap pola yang kompleks pada data berdimensi tinggi. Hal ini sangat relevan dalam analisis ekspresi gen yang memiliki hubungan non-linear antar fitur [13]. Oleh karena itu, *SVM* menjadi salah satu model yang banyak digunakan dalam penelitian bioinformatika.

## 2.7 Multi-Layer Perceptron

*Multi-Layer Perceptron (MLP)* merupakan jaringan saraf tiruan bertipe *feed-forward* yang terdiri dari lapisan input, lapisan tersembunyi, dan lapisan output. Setiap lapisan saling terhubung melalui bobot yang dipelajari selama proses pelatihan. Struktur ini memungkinkan *MLP* memodelkan hubungan yang kompleks antar fitur.

Penggunaan fungsi aktivasi non-linear pada lapisan tersembunyi memungkinkan *MLP* menangkap pola non-linear dalam data. Kemampuan ini menjadikan *MLP* cocok untuk data ekspresi gen yang bersifat kompleks. Dengan konfigurasi yang tepat, *MLP* dapat menghasilkan performa klasifikasi yang kompetitif [14].

## 2.8 Ensemble Learning

*Ensemble learning* merupakan pendekatan yang mengkombinasikan beberapa model untuk meningkatkan stabilitas dan akurasi prediksi. Metode ini bekerja dengan menggabungkan hasil prediksi dari berbagai algoritma pembelajaran. Dengan demikian, kelemahan dari satu model dapat ditutupi oleh model lainnya.

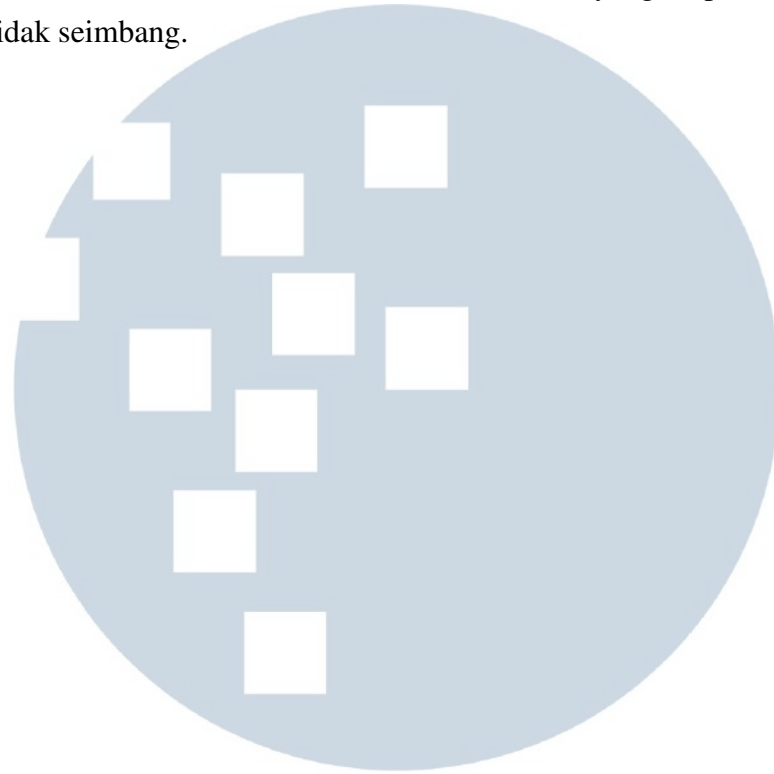
Pada penelitian ini digunakan metode *soft voting*, di mana prediksi akhir ditentukan berdasarkan rata-rata probabilitas dari seluruh model. Pendekatan ini memungkinkan pemanfaatan informasi probabilistik secara lebih optimal. *Ensemble learning* terbukti mampu meningkatkan performa model dibandingkan penggunaan satu model tunggal [15].

## 2.9 Metrik Evaluasi Model

Evaluasi performa model klasifikasi pada penelitian ini dilakukan menggunakan beberapa metrik, yaitu *accuracy*, *precision*, *recall*, *F1-score*, serta *Area Under Curve (AUC)*. Penggunaan berbagai metrik ini bertujuan untuk



memberikan gambaran yang lebih komprehensif mengenai kinerja model. Hal ini penting terutama dalam konteks data klasifikasi dua kelas yang berpotensi memiliki distribusi tidak seimbang.



UMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA