

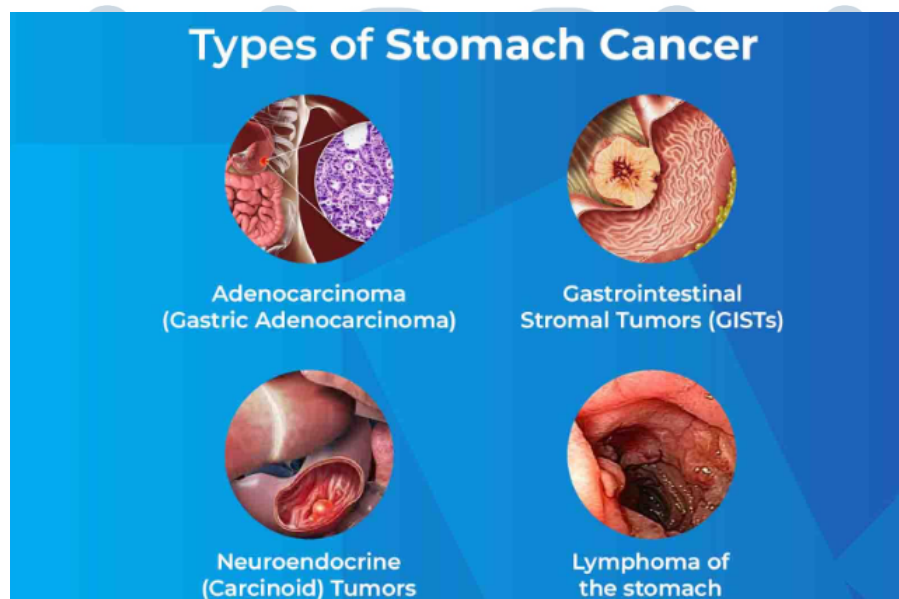
BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker lambung (*gastric cancer*) merupakan salah satu penyebab utama kematian akibat kanker di dunia dan masih menjadi tantangan besar dalam manajemen klinis global. Penyakit ini terjadi akibat pertumbuhan sel abnormal pada jaringan lambung yang berkembang secara tidak terkendali sehingga dapat menginvasi jaringan di sekitarnya maupun menyebar ke organ lain melalui proses metastasis. Menurut data GLOBOCAN 2020, kanker lambung menempati urutan kelima sebagai kanker dengan insidensi tertinggi dan urutan keempat sebagai penyebab kematian akibat kanker di seluruh dunia [1]. Tingginya angka mortalitas tersebut disebabkan oleh heterogenitas biologis tumor, keterlambatan diagnosis, serta perbedaan respons pasien terhadap terapi yang diberikan.

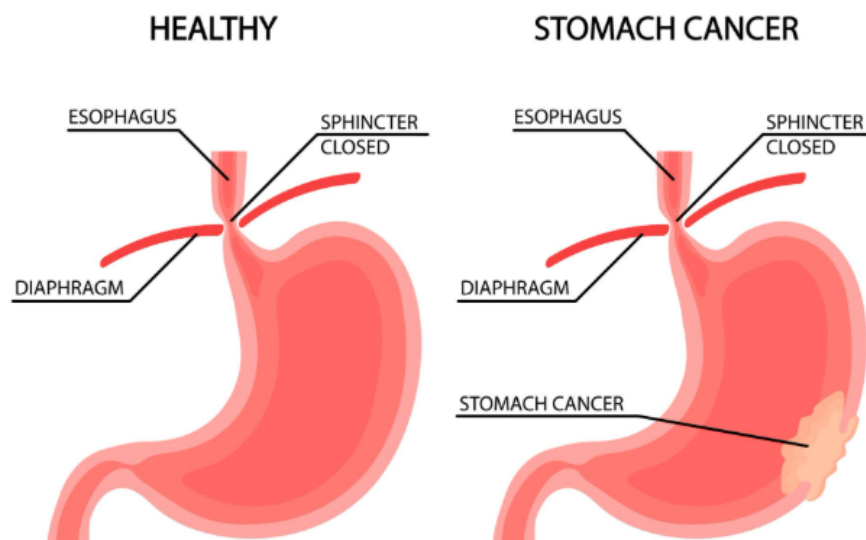
Kanker lambung bukan merupakan satu jenis penyakit yang homogen, melainkan terdiri atas beberapa jenis berdasarkan asal jaringan penyusunnya. Secara histopatologi, kanker lambung dapat dibedakan menjadi adenokarsinoma lambung (*gastric adenocarcinoma*), gastrointestinal stromal tumor (GIST), limfoma lambung, dan tumor neuroendokrin. Di antara berbagai jenis tersebut, adenokarsinoma lambung merupakan tipe yang paling dominan dengan proporsi sekitar 90–95% dari seluruh kasus kanker lambung, sedangkan GIST, limfoma lambung, dan tumor neuroendokrin hanya menyumbang sebagian kecil kasus. Oleh karena itu, sebagian besar penelitian mengenai kanker lambung berfokus pada adenokarsinoma karena merupakan tipe dengan insidensi tertinggi dan telah banyak dikarakterisasi baik secara histopatologi maupun molekuler [?].



Gambar 1.1. Jenis-jenis kanker lambung berdasarkan klasifikasi histopatologi.

Secara anatomi, lambung normal memiliki lapisan mukosa yang tersusun secara teratur

dan berfungsi dalam proses pencernaan makanan. Pada kondisi kanker lambung, sel-sel epitel mengalami proliferasi yang tidak terkendali sehingga membentuk massa tumor yang dapat merusak struktur mukosa, menginvasi lapisan dinding lambung, hingga menyebar ke jaringan dan organ lain melalui proses metastasis. Perubahan morfologi tersebut menyebabkan gangguan fungsi lambung dan sering kali tidak menimbulkan gejala yang spesifik pada stadium awal sehingga banyak kasus baru terdiagnosis ketika penyakit telah memasuki stadium lanjut.



Gambar 1.2. Perbandingan anatomi lambung normal dan lambung yang mengalami kanker.

Secara histopatologi, sekitar 90–95% kasus kanker lambung merupakan adenokarsinoma lambung (*gastric adenocarcinoma*), sedangkan sisanya terdiri atas jenis kanker yang lebih jarang seperti gastrointestinal stromal tumor (GIST), limfoma lambung, dan tumor neuroendokrin. Adenokarsinoma lambung berasal dari sel epitel mukosa lambung dan merupakan bentuk kanker lambung yang paling banyak dipelajari karena memiliki angka kejadian tertinggi. Berdasarkan klasifikasi Lauren, adenokarsinoma lambung dibedakan menjadi dua tipe utama, yaitu tipe intestinal dan tipe diffuse. Tipe intestinal ditandai oleh pembentukan struktur kelenjar yang menyerupai epitel usus, umumnya berkembang melalui tahapan lesi prakanker akibat infeksi *Helicobacter pylori*, dan memiliki prognosis yang relatif lebih baik. Sebaliknya, tipe diffuse ditandai oleh penyebaran sel tumor secara infiltratif tanpa membentuk struktur kelenjar yang jelas, sering ditemukan pada pasien usia muda, serta memiliki sifat invasif dan prognosis yang lebih buruk. Selain kedua tipe tersebut, terdapat tipe mixed yang menunjukkan karakteristik gabungan antara tipe intestinal dan diffuse [2, 3].

Meskipun klasifikasi histopatologi telah lama digunakan dalam praktik klinis, pendekatan tersebut belum mampu menjelaskan sepenuhnya heterogenitas biologis kanker lambung maupun variasi respons pasien terhadap terapi. Pasien dengan tipe histopatologi yang sama dapat memiliki karakteristik molekuler yang berbeda sehingga menunjukkan prognosis dan efektivitas terapi yang

tidak selalu sama. Oleh karena itu, perkembangan teknologi genomik mendorong lahirnya klasifikasi berbasis molekuler yang mampu memberikan pemahaman lebih komprehensif mengenai mekanisme biologis kanker lambung.

Perkembangan teknologi genomik telah menghasilkan klasifikasi molekuler kanker lambung yang diperkenalkan oleh The Cancer Genome Atlas (TCGA). TCGA mengelompokkan kanker lambung menjadi empat sub tipe molekuler utama, yaitu Epstein-Barr Virus (EBV)-positive, Microsatellite Instability (MSI), Genomically Stable (GS), dan Chromosomal Instability (CIN) [4]. Sub tipe EBV-positive ditandai oleh infeksi virus Epstein-Barr, hipermetilasi DNA, serta tingginya ekspresi gen yang berkaitan dengan respons imun. Sub tipe Genomically Stable (GS) memiliki tingkat perubahan kromosom yang rendah dan umumnya berasosiasi dengan kanker lambung tipe diffuse. Sementara itu, sub tipe Chromosomal Instability (CIN) ditandai oleh tingginya instabilitas kromosom, amplifikasi gen, serta aktivasi berbagai jalur pensinyalan yang berperan dalam proliferasi sel tumor. Adapun sub tipe Microsatellite Instability (MSI) ditandai oleh gangguan pada sistem DNA mismatch repair (MMR) yang menyebabkan akumulasi mutasi pada daerah mikrosatelit genom [5].

Di antara keempat sub tipe tersebut, Microsatellite Instability-High (MSI-H) menjadi salah satu sub tipe yang paling banyak mendapat perhatian karena memiliki karakteristik biologis yang berbeda dibandingkan dengan microsatellite stable (MSS) atau non-MSI-H. MSI-H umumnya dikaitkan dengan tingginya *tumor mutation burden* (TMB), infiltrasi sel imun yang lebih kuat, serta respons yang lebih baik terhadap terapi imun berbasis *immune checkpoint inhibitor* [6, 7]. Oleh karena itu, identifikasi status MSI memiliki nilai prognostik dan prediktif yang sangat penting dalam menentukan strategi terapi yang tepat bagi pasien kanker lambung.

Perkembangan kecerdasan buatan (*Artificial Intelligence/AI*) telah mendorong berbagai pendekatan untuk mengklasifikasi MSI-H dan non-MSI-H secara otomatis. Salah satu penelitian penting dilakukan oleh Kather *et al.* yang menunjukkan bahwa model *deep learning* mampu memprediksi status MSI secara langsung dari citra histopatologi hematoxylin-eosin (H&E) tanpa memerlukan pemeriksaan molekuler tambahan [8]. Selanjutnya, pendekatan berbasis *Deep Gaussian Process* juga dikembangkan untuk meningkatkan estimasi ketidakpastian model dalam prediksi MSI dan respons imunoterapi dari citra histopatologi digital [9].

Selain pendekatan histopatologi, metode radiomics berbasis *computed tomography* (CT) telah banyak digunakan untuk mendeteksi MSI secara non-invasif. Penelitian menunjukkan bahwa fitur radiomik yang diekstraksi dari citra CT mampu memprediksi status MSI dan potensi respons terhadap imunoterapi pada pasien kanker lambung [10, 11]. Meskipun menunjukkan performa yang menjanjikan, pendekatan berbasis citra tersebut tidak secara langsung mengeksplorasi mekanisme biologis yang mendasari status MSI pada tingkat molekuler.

Seiring berkembangnya teknologi *high-throughput sequencing*, pendekatan berbasis data molekuler seperti RNA sequencing (RNA-seq) dan DNA methylation mulai digunakan untuk mendeteksi status MSI. RNA-seq memungkinkan pengukuran ekspresi ribuan gen secara simultan sehingga dapat memberikan gambaran aktivitas biologis tumor secara komprehensif [12]. Salah satu penelitian terbaru, MSI-XGNN, mengintegrasikan data transkriptom dan metilasi menggunakan pendekatan *Graph Neural Network* untuk meningkatkan akurasi deteksi MSI sekaligus mengidentifikasi biomarker molekuler yang relevan [13]. Namun demikian, sebagian besar penelitian tersebut berfokus pada pengembangan model klasifikasi MSI berbasis data molekuler tanpa mempertimbangkan kontribusi faktor klinikal pasien.

Berdasarkan telaah literatur, terdapat beberapa kesenjangan penelitian (*research gap*) yang masih terbuka. Pertama, sebagian besar penelitian berbasis AI untuk deteksi MSI pada kanker lambung berfokus pada data citra histopatologi atau radiologi, sementara pendekatan berbasis ekspresi gen (*transcriptome-based classification*) masih relatif terbatas. Kedua, studi yang menggunakan data RNA-seq umumnya hanya memanfaatkan fitur molekuler tanpa mengintegrasikan variabel klinikal seperti usia, jenis kelamin, stadium tumor, dan karakteristik patologi pasien. Ketiga, masih terbatas penelitian yang secara sistematis membandingkan performa model berbasis *gene expression* saja, data klinikal saja, dan model integratif yang menggabungkan kedua jenis data tersebut dalam klasifikasi MSI-H dan non-MSI-H pada kanker lambung.

Pendekatan berbasis ekspresi gen memiliki keunggulan karena mampu memberikan interpretasi biologis yang lebih mendalam dibandingkan metode berbasis citra. Profil ekspresi gen mencerminkan aktivitas jalur molekuler yang berkaitan langsung dengan mekanisme *DNA mismatch repair* (MMR), regulasi siklus sel, proliferasi tumor, serta respons imun antitumor. Dengan demikian, selain berfungsi sebagai fitur prediktif, data transkriptom juga dapat digunakan untuk mengidentifikasi biomarker potensial yang berhubungan dengan status MSI [13].

Di sisi lain, beberapa kajian menunjukkan bahwa karakteristik klinis pasien memiliki keterkaitan dengan distribusi subtype molekuler kanker lambung [4, 14]. Oleh karena itu, integrasi data molekuler dan data klinikal berpotensi menghasilkan model klasifikasi yang lebih akurat dan lebih representatif terhadap kondisi pasien di dunia nyata. Namun hingga saat ini, penelitian yang secara khusus mengevaluasi kontribusi variabel klinikal terhadap performa model klasifikasi MSI berbasis ekspresi gen pada kanker lambung masih sangat terbatas.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi status MSI-H dan non-MSI-H pada kanker lambung menggunakan data ekspresi gen yang diintegrasikan dengan variabel klinikal pasien. Selain mengevaluasi performa model klasifikasi menggunakan beberapa algoritma *machine learning*, penelitian ini juga bertujuan mengidentifikasi biomarker gen yang berkontribusi terhadap klasifikasi MSI melalui proses seleksi fitur dan interpretasi model. Hasil penelitian ini diharapkan dapat memberikan pemahaman biologis yang lebih mendalam mengenai karakteristik molekuler kanker lambung serta mendukung pengembangan pendekatan *precision oncology* yang lebih efektif dalam membantu pengambilan keputusan klinis bagi pasien kanker lambung.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana membangun model klasifikasi untuk membedakan pasien kanker lambung dengan status MSI-H dan non-MSI-H menggunakan data *gene expression*?
2. Bagaimana pengaruh metode seleksi fitur, yaitu Limma, LASSO, ANOVA, dan Recursive Feature Elimination (RFE), terhadap performa model klasifikasi MSI-H dan non-MSI-H?
3. Bagaimana pengaruh penggunaan top 50 gen hasil seleksi Differentially Expressed Genes (DEG) terhadap kinerja model klasifikasi MSI-H dan non-MSI-H?

4. Algoritma machine learning manakah yang memberikan performa terbaik dalam mengklasifikasikan MSI-H dan non-MSI-H setelah dilakukan penanganan ketidakseimbangan kelas menggunakan metode SMOTE?
5. Gen-gen apa saja yang berkontribusi paling signifikan dalam membedakan MSI-H dan non-MSI-H berdasarkan interpretasi model menggunakan metode SHAP?

1.3 Batasan Permasalahan

Agar penelitian ini terarah dan tidak menyimpang dari tujuan yang telah ditetapkan, maka ditetapkan beberapa batasan permasalahan sebagai berikut:

1. Penelitian ini difokuskan pada klasifikasi status Microsatellite Instability, yaitu MSI-High (MSI-H) dan non-MSI-H pada pasien kanker lambung.
2. Data yang digunakan berupa data *gene expression* (RNA-seq) dan data klinikal pasien yang tersedia pada dataset publik, seperti TCGA-STAD atau dataset sejenis.
3. Metode seleksi fitur yang digunakan dibatasi pada Limma, LASSO, ANOVA, dan Recursive Feature Elimination (RFE), sedangkan algoritma klasifikasi yang digunakan meliputi Support Vector Machine (SVM), Random Forest, dan XGBoost.
4. Penanganan ketidakseimbangan kelas dilakukan menggunakan metode SMOTE, sedangkan evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, F1-score, AUC, dan cross-validation.
5. Penelitian ini tidak mencakup analisis citra histopatologi, radiologi, data multi-omics lainnya, analisis survival, maupun validasi klinis prospektif.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disusun, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model machine learning untuk mengklasifikasi status MSI-High (MSI-H) dan non-MSI-H pada pasien kanker lambung menggunakan data gene expression.
2. Menganalisis pengaruh integrasi variabel klinikal (seperti usia, jenis kelamin, dan stadium tumor) terhadap peningkatan performa model klasifikasi dibandingkan dengan model yang hanya menggunakan data gene expression.
3. Membandingkan performa beberapa algoritma machine learning dalam mengklasifikasi MSI-H dan non-MSI-H berdasarkan metrik evaluasi seperti accuracy, precision, recall, F1-score, dan Area Under the Curve (AUC).
4. Mengidentifikasi gen-gen yang memiliki kontribusi paling signifikan dalam membedakan MSI-H dan non-MSI-H melalui analisis feature importance atau metode interpretabilitas model.
5. Mengevaluasi kestabilan dan generalisasi model melalui proses validasi silang (cross-validation) dan/atau validasi menggunakan dataset eksternal.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis sebagai berikut:

1.5.1 Manfaat Teoritis

1. Memberikan kontribusi ilmiah dalam pengembangan metode klasifikasi berbasis data transcriptome untuk membedakan status MSI-High (MSI-H) dan non-MSI-H pada kanker lambung.
2. Menambah kajian mengenai integrasi data molekuler dan variabel klinikal dalam model machine learning pada bidang bioinformatika dan computational oncology.
3. Menjadi referensi bagi penelitian selanjutnya yang berkaitan dengan pengembangan biomarker berbasis gene expression untuk klasifikasi molekuler kanker.

1.5.2 Manfaat Praktis

1. Memberikan pendekatan alternatif dalam mendukung identifikasi status MSI secara komputasional berbasis data genomik, yang berpotensi melengkapi metode pemeriksaan konvensional.
2. Membantu dalam pengembangan sistem pendukung keputusan klinis (clinical decision support system) berbasis data molekuler dan klinikal.
3. Memberikan informasi mengenai gen-gen yang berperan penting dalam membedakan MSI-H dan non-MSI-H, yang berpotensi menjadi kandidat biomarker penelitian lanjutan.
4. Menjadi dasar pengembangan model prediktif yang lebih terintegrasi dan presisi dalam pendekatan precision medicine pada kanker lambung.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
Bab ini berisi pendahuluan yang meliputi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.
- Bab 2 LANDASAN TEORI
Bab ini memuat tinjauan pustaka dan dasar teori yang mendukung penelitian, meliputi konsep Microsatellite Instability (MSI), karakteristik MSI-H dan non-MSI-H pada kanker lambung, gene expression (RNA-seq), variabel klinikal, serta metode machine learning yang digunakan untuk klasifikasi.

- Bab 3 METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian yang mencakup sumber dan karakteristik dataset, tahapan preprocessing data, seleksi fitur, perancangan model machine learning, skema pelatihan dan validasi, serta metrik evaluasi yang digunakan dalam mengukur performa model klasifikasi.

- Bab 4 HASIL DAN DISKUSI

Bab ini menyajikan hasil dan pembahasan penelitian, meliputi hasil eksperimen model klasifikasi, perbandingan performa antar model, analisis kontribusi fitur (feature importance), serta interpretasi hasil dalam konteks biologis dan klinis.

- Bab 5 KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan serta saran untuk pengembangan penelitian selanjutnya.

