BAB 2
LANDASAN TEORI

Bab ini membahas landasan teori dan kajian pustaka yang mendukung penelitian, meliputi konsep dasar kanker lambung, peran microRNA sebagai *biomarker* molekuler, permasalahan data berdimensi tinggi, seleksi fitur berbasis *mutual information*, teori *fuzzy*, *Fuzzy Mutual Information*, serta algoritma *machine learning* yang digunakan dalam prediksi stadium kanker lambung.

2.1 MicroRNA sebagai Biomarker Molekuler Kanker Lambung

MicroRNA (miRNA) merupakan molekul RNA non-pengkode berukuran kecil, umumnya terdiri dari sekitar 18–25 nukleotida, yang berperan dalam regulasi ekspresi gen pada tingkat post-transkripsi melalui mekanisme degradasi mRNA atau penghambatan translasi. Dalam konteks kanker, miRNA diketahui terlibat dalam berbagai proses biologis penting, termasuk proliferasi sel, diferensiasi, apoptosis, dan metastasis. Perubahan pola ekspresi miRNA dapat mencerminkan kondisi patologis tertentu, sehingga miRNA banyak diteliti sebagai kandidat biomarker molekuler untuk diagnosis, prognosis, dan prediksi stadium penyakit kanker [8].

Pada kanker lambung, sejumlah penelitian menunjukkan bahwa perbedaan ekspresi miRNA berkorelasi erat dengan perkembangan stadium penyakit, termasuk peralihan dari stadium awal ke stadium lanjut. miRNA tertentu dapat berperan sebagai onkomiR maupun *tumor suppressor*, bergantung pada gen target yang diregulasinya. Oleh karena itu, analisis ekspresi miRNA memberikan informasi molekuler yang lebih spesifik dibandingkan parameter klinis konvensional, terutama dalam mendukung proses klasifikasi stadium kanker berbasis data genomik [3].

Meskipun demikian, data ekspresi miRNA memiliki karakteristik berdimensi tinggi dengan jumlah fitur yang jauh melebihi jumlah sampel, khususnya pada dataset publik seperti The Cancer Genome Atlas (TCGA). Karakteristik ini menimbulkan tantangan analitik berupa redundansi fitur, korelasi antar miRNA, serta keberadaan noise, yang dapat menurunkan kinerja model prediksi apabila tidak disertai dengan strategi seleksi fitur yang tepat [6].

2.2 Stage Stadium Kanker

Stadium kanker merupakan indikator klinis penting yang digunakan untuk menggambarkan tingkat perkembangan penyakit, serta menjadi dasar dalam penentuan strategi terapi dan estimasi prognosis pasien. Pada kanker lambung, sistem klasifikasi stadium yang umum digunakan adalah sistem TNM (Tumor, Node, Metastasis) yang dikembangkan oleh *American Joint Committee on Cancer* (AJCC). Komponen T (*pathologic T-stage*) merepresentasikan tingkat invasi tumor primer ke lapisan dinding lambung dan jaringan sekitarnya, sehingga memiliki relevansi langsung terhadap progresivitas penyakit [9].

Dalam penelitian ini, klasifikasi stadium kanker lambung difokuskan pada *pathologic T-stage*, yang kemudian dikelompokkan menjadi dua kategori, yaitu stadium awal (T1 dan T2) dan stadium lanjut (T3 dan T4). Pengelompokan biner ini dilakukan untuk menyederhanakan permasalahan klasifikasi sekaligus mencerminkan perbedaan klinis yang signifikan antara kedua

kelompok tersebut. Stadium awal umumnya menunjukkan keterbatasan invasi tumor, sedangkan stadium lanjut mengindikasikan penetrasi yang lebih dalam serta peningkatan risiko penyebaran penyakit [3].

Pendekatan pengelompokan stadium menjadi dua kelas juga sejalan dengan tujuan klinis, yaitu meningkatkan sensitivitas deteksi stadium lanjut yang memerlukan penanganan lebih agresif. Dengan demikian, klasifikasi stadium berbasis T-*stage* tidak hanya relevan secara klinis, tetapi juga sesuai untuk dikembangkan sebagai permasalahan klasifikasi berbasis data genomik menggunakan pendekatan *machine learning*.

2.3 Seleksi Fitur pada Data Berdimensi Tinggi

Seleksi fitur merupakan tahapan penting dalam analisis data genomik untuk mengidentifikasi subset fitur yang paling relevan terhadap variabel target. Pada data miRNA, seleksi fitur bertujuan untuk mengurangi kompleksitas model, meningkatkan generalisasi, serta meminimalkan risiko *overfitting* akibat rasio fitur terhadap sampel yang sangat tinggi. Pendekatan seleksi fitur umumnya diklasifikasikan menjadi metode *filter*, *wrapper*, dan *embedded*, dengan metode filter sering dipilih pada data genomik karena efisiensinya dan independensinya terhadap model klasifikasi tertentu [6].

Metode filter konvensional berbasis statistik atau *mutual information* klasik mengukur hubungan antara fitur dan label kelas berdasarkan distribusi probabilitas yang bersifat tegas (*crisp*). Namun, pendekatan tersebut memiliki keterbatasan dalam menangkap ketidakpastian dan hubungan nonlinier yang kompleks pada data biologis, terutama ketika data bersifat kontinu dan mengandung variabilitas tinggi [10]. Oleh karena itu, diperlukan pendekatan seleksi fitur yang lebih adaptif dan mampu merepresentasikan ketidakpastian intrinsik pada data biomedis. Integrasi teori *fuzzy* dalam proses seleksi fitur menjadi salah satu solusi yang berkembang untuk mengatasi keterbatasan metode konvensional, khususnya pada analisis data genomik berdimensi tinggi.

2.4 Teori Fuzzy dalam Analisis Data Biomedis

Teori *fuzzy* diperkenalkan oleh Zadeh sebagai kerangka matematis untuk merepresentasikan ketidakpastian, ambiguitas, dan ketidaktegasan yang tidak dapat dimodelkan secara efektif menggunakan logika biner klasik [11]. Berbeda dengan pendekatan *crisp*, teori *fuzzy* memungkinkan suatu elemen memiliki derajat keanggotaan kontinu dalam interval [0,1], sehingga lebih fleksibel dalam merepresentasikan fenomena dunia nyata.

Dalam konteks data biomedis, khususnya data ekspresi gen dan miRNA, ketidakpastian sering muncul akibat variabilitas biologis, noise eksperimen, serta perbedaan kondisi pengukuran. Pendekatan *fuzzy* dinilai lebih sesuai untuk menangkap karakteristik tersebut karena mampu memodelkan hubungan gradual antar nilai fitur, alih-alih melakukan diskretisasi keras yang berpotensi menghilangkan informasi penting [12].

Penerapan teori *fuzzy* dalam analisis data genomik telah banyak digunakan pada berbagai tahap pemrosesan data, termasuk klasterisasi, klasifikasi, dan seleksi fitur. Salah satu pendekatan yang berkembang adalah penggabungan teori *fuzzy* dengan konsep informasi untuk mengukur ketergantungan antar variabel secara lebih adaptif dan robust terhadap ketidakpastian data.

2.5 Fuzzy Mutual Information dan True Fuzzy Mutual Information

Fuzzy Mutual Information (FMI) merupakan pengembangan dari konsep *mutual information* klasik yang mengintegrasikan teori *fuzzy* untuk mengukur tingkat ketergantungan antara fitur dan label kelas. Berbeda dengan mutual information konvensional yang bergantung pada estimasi probabilitas diskret, FMI memanfaatkan relasi *fuzzy* untuk merepresentasikan kedekatan antar sampel secara kontinu, sehingga lebih sesuai untuk data numerik berdimensi tinggi [7].

Dalam penelitian ini, digunakan pendekatan *True Fuzzy Mutual Information* yang dikembangkan oleh Salem et al. [13]. Pendekatan ini mendefinisikan relasi ekuivalensi *fuzzy* antar sampel menggunakan fungsi keanggotaan berbasis jarak, kemudian menghitung *fuzzy entropy* dan *fuzzy joint entropy* untuk mengukur ketergantungan informasi secara lebih akurat. Nilai *mutual information fuzzy* dihitung sebagai kombinasi antara entropi fitur, entropi label, dan entropi gabungan, sehingga mampu merepresentasikan hubungan nonlinier antara miRNA dan stadium kanker.

Pendekatan *True Fuzzy Mutual Information* dinilai unggul dibandingkan metode seleksi fitur konvensional karena mampu mempertahankan informasi kontinu, mengurangi sensitivitas terhadap noise, serta memberikan evaluasi relevansi fitur yang lebih stabil pada data genomik berdimensi tinggi.

2.6 Penanganan Ketidakseimbangan Data dengan SMOTETomek

Ketidakseimbangan kelas merupakan permasalahan umum pada data klinis dan genomik, termasuk dalam klasifikasi stadium kanker, di mana jumlah sampel pada satu kelas sering kali lebih dominan dibandingkan kelas lainnya. Kondisi ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas dan menurunkan sensitivitas terhadap kelas minoritas, yang dalam konteks klinis justru sering memiliki tingkat kepentingan yang lebih tinggi. Oleh karena itu, diperlukan strategi penyeimbangan data yang tidak hanya meningkatkan jumlah sampel kelas minoritas, tetapi juga memperbaiki kualitas batas keputusan antar kelas.

SMOTETomek merupakan metode *hybrid resampling* yang mengombinasikan pendekatan *oversampling* dan *undersampling*. Tahap pertama menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menghasilkan sampel sintetis pada kelas minoritas dengan cara melakukan interpolasi linier antara suatu sampel minoritas dan tetangga terdekatnya dalam ruang fitur. Proses ini bertujuan untuk memperkaya representasi kelas minoritas tanpa melakukan duplikasi data secara langsung, sehingga mengurangi risiko *overfitting* dibandingkan metode *random oversampling* [14]. Secara matematis, sampel sintetis dihasilkan berdasarkan kombinasi linier antara dua sampel minoritas yang berdekatan, sehingga tetap berada dalam distribusi asli kelas tersebut.

Setelah proses SMOTE, tahap kedua dilakukan menggunakan Tomek Links, yaitu pasangan dua sampel dari kelas yang berbeda yang saling menjadi tetangga terdekat satu sama lain. Keberadaan Tomek Links mengindikasikan adanya tumpang tindih atau ambiguitas pada batas keputusan antar kelas. Dalam SMOTETomek, sampel dari pasangan Tomek Links—umumnya berasal dari kelas mayoritas—dihapus untuk memperjelas pemisahan antar kelas dan mengurangi noise di sekitar boundary [15]. Dengan demikian, SMOTETomek tidak hanya menyeimbangkan distribusi kelas, tetapi juga meningkatkan separabilitas data dengan membersihkan area yang

berpotensi menyebabkan kesalahan klasifikasi.

Pemilihan SMOTETomek [16] dalam penelitian ini didasarkan pada karakteristik data ekspresi miRNA yang berdimensi tinggi dan memiliki potensi tumpang tindih antar kelas stadium. Metode oversampling murni seperti SMOTE atau Borderline-SMOTE cenderung menambah sampel sintetis di sekitar boundary tanpa mekanisme eksplisit untuk menghapus sampel ambigu, sehingga berisiko memperkuat noise pada data biologis. Sebaliknya, SMOTETomek mengombinasikan keunggulan pembangkitan sampel sintetis dengan pembersihan batas kelas, sehingga dinilai lebih sesuai untuk meningkatkan stabilitas model klasifikasi pada data genomik dan mendukung deteksi kelas minoritas secara lebih akurat.

2.7 Machine Learning untuk Klasifikasi Stadium Kanker

Algoritma *supervised machine learning* telah banyak diterapkan dalam klasifikasi kanker berbasis data genomik karena kemampuannya dalam mempelajari pola kompleks dari data berdimensi tinggi. Dalam penelitian ini, digunakan tiga algoritma klasifikasi, yaitu *Support Vector Machine* (SVM), *Random Forest* (RF), dan *K-Nearest Neighbors* (KNN), yang masing-masing memiliki karakteristik dan keunggulan berbeda.

SVM dikenal efektif dalam menangani data berdimensi tinggi dengan memaksimalkan margin pemisah antar kelas, terutama ketika dikombinasikan dengan kernel nonlinier seperti *Radial Basis Function* (RBF). *Random Forest* merupakan metode *ensemble* berbasis pohon keputusan yang robust terhadap noise dan mampu menangani interaksi fitur yang kompleks. Sementara itu, KNN memanfaatkan kedekatan antar sampel dalam ruang fitur dan bersifat non-parametrik, sehingga dapat menangkap pola lokal dalam data [17].

Penggunaan beberapa algoritma klasifikasi bertujuan untuk mengevaluasi konsistensi kinerja subset fitur terpilih lintas model, serta memastikan bahwa seleksi fitur yang diusulkan tidak bergantung pada satu model tertentu.

2.8 Evaluasi Model dan Validasi

Evaluasi dan validasi model merupakan tahapan penting dalam memastikan bahwa model klasifikasi yang dikembangkan memiliki kinerja yang andal dan mampu melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Pada penelitian ini, evaluasi model dilakukan menggunakan skema *Stratified K-Fold Cross-Validation*, di mana data dibagi menjadi sejumlah lipatan (*fold*) dengan proporsi kelas yang tetap terjaga pada setiap lipatan. Pendekatan ini bertujuan untuk meminimalkan bias evaluasi serta menghasilkan estimasi kinerja model yang lebih stabil [18].

Kinerja model dievaluasi menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall* atau sensitivitas, F1-score, serta *Area Under the Receiver Operating Characteristic Curve* (AUC). Penggunaan beberapa metrik dilakukan untuk memberikan gambaran kinerja yang lebih komprehensif, terutama pada kondisi data yang tidak seimbang. Sensitivitas menjadi metrik penting dalam penelitian ini karena berkaitan langsung dengan kemampuan model dalam mendeteksi stadium lanjut kanker, yang memiliki implikasi klinis lebih besar [19].

Selain evaluasi numerik, analisis *confusion matrix* digunakan untuk mengamati distribusi prediksi benar dan salah pada masing-masing kelas. Normalisasi *confusion matrix* juga dilakukan

untuk mempermudah interpretasi proporsi kesalahan klasifikasi. Kombinasi antara evaluasi berbasis metrik kuantitatif dan analisis *confusion matrix* diharapkan mampu memberikan pemahaman yang lebih menyeluruh terhadap performa model serta keterbatasan pendekatan yang diusulkan.