# BAB 1
# PENDAHULUAN

## 1.1 Latar Belakang Masalah

Kanker lambung (*gastric cancer*) masih menjadi salah satu tantangan utama dalam bidang kesehatan global, khususnya di negara-negara berkembang karena tingginya angka mortalitas yang ditimbulkannya [1]. Secara global, penyakit ini mencatat lebih dari satu juta kasus baru dan sekitar 783.000 kematian pada tahun 2018, sehingga menempatkannya sebagai kanker kelima yang paling sering didiagnosis dan penyebab kematian akibat kanker ketiga di dunia [2]. Tingginya angka kematian pada kanker lambung sebagian besar disebabkan oleh keterlambatan diagnosis, di mana mayoritas pasien baru terdeteksi pada stadium lanjut. Perbedaan karakteristik biologis antara stadium awal dan stadium lanjut memiliki peran penting dalam menentukan strategi terapi serta peluang keberhasilan pengobatan. Oleh karena itu, diperlukan pendekatan analitik yang mampu mendukung proses klasifikasi stadium kanker secara lebih akurat dan berbasis data molekuler [3].

Perkembangan teknologi *high-throughput sequencing* telah memungkinkan pengukuran ekspresi microRNA (miRNA) secara komprehensif dalam skala besar. miRNA merupakan molekul RNA non-pengkode berukuran kecil yang berperan sebagai regulator ekspresi gen pada tingkat post-transkripsi [4]. Sejumlah penelitian menunjukkan bahwa perubahan pola ekspresi miRNA memiliki keterkaitan erat dengan proses karsinogenesis dan progresi kanker, termasuk kanker lambung, sehingga miRNA berpotensi dimanfaatkan sebagai *biomarker* molekuler untuk memprediksi stadium penyakit [5]. Meskipun demikian, data ekspresi miRNA umumnya memiliki karakteristik berdimensi tinggi (*high-dimensional*), dengan jumlah fitur yang jauh lebih besar dibandingkan jumlah sampel, khususnya pada dataset publik seperti The Cancer Genome Atlas (TCGA). Kondisi ini dapat menurunkan kinerja model prediksi akibat keberadaan fitur yang tidak relevan, redundan, atau bersifat noise apabila tidak ditangani dengan pendekatan analitik yang tepat [6].

Pendekatan *machine learning* telah banyak diterapkan dalam penelitian kanker untuk melakukan klasifikasi dan prediksi berbasis data genomik. Namun, kinerja model *machine learning* sangat bergantung pada kualitas fitur yang digunakan sebagai masukan. Oleh karena itu, seleksi fitur menjadi tahap yang krusial untuk mengidentifikasi subset miRNA yang paling relevan terhadap stadium kanker, sekaligus mengurangi kompleksitas model dan risiko *overfitting*. Metode seleksi fitur konvensional berbasis statistik atau *mutual information* klasik sering kali belum optimal dalam menangkap ketidakpastian serta hubungan nonlinier yang kompleks pada data biologis.

Teori *fuzzy* menawarkan kerangka matematis yang lebih fleksibel dalam merepresentasikan ketidakpastian dan ambiguitas yang umum dijumpai pada data biomedis. Salah satu pendekatan yang berkembang adalah seleksi fitur berbasis *Fuzzy Mutual Information* [7], yang mengintegrasikan konsep *mutual information* dengan teori *fuzzy* untuk mengukur tingkat ketergantungan antara fitur dan label kelas secara lebih adaptif. Pendekatan ini memungkinkan evaluasi relevansi fitur dengan mempertimbangkan sifat kontinu dan tidak pasti dari data ekspresi miRNA, sehingga dinilai lebih sesuai untuk analisis data genomik berdimensi tinggi dibandingkan metode seleksi fitur konvensional.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan penerapan metode seleksi fitur berbasis *Fuzzy Mutual Information* pada data miRNA TCGA untuk klasifikasi stadium kanker lambung. Subset miRNA terpilih selanjutnya digunakan sebagai masukan bagi beberapa model *machine learning*, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Random Forest* (RF). Untuk mengatasi permasalahan ketidakseimbangan kelas, diterapkan pula teknik *resampling*. Evaluasi kinerja model dilakukan menggunakan skema *Stratified K-Fold Cross-Validation* dan metrik evaluasi yang sesuai, sehingga pendekatan yang diusulkan diharapkan mampu menghasilkan kinerja prediksi yang optimal serta berkontribusi dalam pengembangan sistem pendukung keputusan berbasis data genomik.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, diperlukan perumusan masalah yang jelas untuk mengarahkan penelitian agar tetap fokus pada permasalahan utama yang ingin diselesaikan, khususnya terkait pemanfaatan data miRNA berdimensi tinggi untuk prediksi stadium kanker lambung. Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan metode seleksi fitur berbasis *Fuzzy Mutual Information* pada data miRNA TCGA untuk mengidentifikasi fitur yang paling relevan dalam prediksi stadium kanker lambung?
2. Bagaimana kinerja model *Support Vector Machine*, *K-Nearest Neighbors*, dan *Random Forest* dalam memprediksi stadium kanker lambung menggunakan fitur miRNA terpilih?

## 1.3 Batasan Permasalahan

Guna memastikan penelitian berjalan secara terarah serta menjaga kedalaman dan konsistensi analisis, ruang lingkup penelitian ini dibatasi pada beberapa aspek tertentu. Pembatasan tersebut ditetapkan agar metode dan hasil yang diperoleh tetap relevan dengan tujuan penelitian serta dapat dianalisis secara sistematis.

1. Data yang digunakan merupakan data ekspresi miRNA dan data klinis kanker lambung yang bersumber dari The Cancer Genome Atlas (TCGA).
2. Klasifikasi stadium kanker lambung dibatasi pada dua kelas, yaitu stadium awal (T1–T2) dan stadium lanjut (T3–T4) berdasarkan parameter *AJCC pathologic T stage*.
3. Metode seleksi fitur yang digunakan adalah *Fuzzy Mutual Information* berbasis teori entropi *fuzzy* dengan pendekatan relevansi–redundansi.
4. Jumlah fitur miRNA yang dipilih dibatasi pada 15 fitur teratas hasil seleksi.
5. Model *machine learning* yang digunakan terbatas pada *Support Vector Machine*, *K-Nearest Neighbors*, dan *Random Forest*.
6. Evaluasi kinerja model dilakukan menggunakan skema *Stratified K-Fold Cross-Validation* dengan metrik evaluasi ROC-AUC, *F1-score*, *Balanced Accuracy*, serta analisis *confusion matrix* ter-normalisasi.

## 1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan pendekatan prediksi stadium kanker lambung berbasis data miRNA dengan memanfaatkan metode seleksi fitur *fuzzy* dan algoritma *machine learning*. Tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan metode *Fuzzy Mutual Information* untuk melakukan seleksi fitur pada data miRNA berdimensi tinggi.
2. Menganalisis dan membandingkan kinerja model *Support Vector Machine*, *K-Nearest Neighbors*, dan *Random Forest* dalam memprediksi stadium kanker lambung menggunakan fitur miRNA terpilih.

## 1.5 Urgensi Penelitian

Kanker lambung memiliki tingkat mortalitas yang tinggi, di mana ketepatan klasifikasi stadium berperan penting dalam menentukan strategi terapi dan prognosis pasien. Data ekspresi microRNA (miRNA) berpotensi menjadi biomarker molekuler untuk klasifikasi stadium kanker, namun karakteristiknya yang berdimensi tinggi dan mengandung ketidakpastian biologis memerlukan metode analisis yang tepat. Oleh karena itu, penerapan seleksi fitur berbasis Fuzzy Mutual Information menjadi penting untuk mengidentifikasi miRNA yang paling relevan dan meningkatkan kinerja model *machine learning*. Penelitian ini memiliki urgensi ilmiah dan aplikatif karena mendukung pengembangan pendekatan prediksi stadium kanker lambung berbasis data genomik dalam konteks Program MBKM Penelitian.

## 1.6 Luaran Penelitian

Penelitian ini menghasilkan luaran yang bersifat akademik dan aplikatif sebagai bentuk kontribusi kegiatan MBKM Penelitian dalam pengembangan ilmu pengetahuan dan teknologi di bidang bioinformatika. Luaran yang dihasilkan tidak hanya berupa dokumentasi ilmiah, tetapi juga mencakup pengembangan model analisis berbasis data genomik yang dapat dimanfaatkan untuk penelitian lanjutan.

1. Laporan ilmiah: Hasil penelitian ini berpotensi disusun menjadi artikel ilmiah yang membahas penerapan seleksi fitur berbasis Fuzzy Mutual Information pada data ekspresi miRNA untuk klasifikasi stadium kanker lambung. Artikel ini direncanakan untuk dipublikasikan pada jurnal nasional terakreditasi atau jurnal internasional bereputasi di bidang bioinformatika, kecerdasan artifisial, atau kesehatan berbasis data.
2. Model klasifikasi: Penelitian ini menghasilkan model klasifikasi stadium kanker lambung berbasis *machine learning* dengan fitur miRNA terpilih hasil seleksi Fuzzy Mutual Information. Model yang dikembangkan diharapkan dapat menjadi dasar pengembangan sistem pendukung keputusan berbasis data genomik serta referensi metodologis bagi penelitian selanjutnya dalam analisis kanker berbasis kecerdasan artifisial.

## 1.7 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan kontribusi baik secara teoritis maupun praktis dalam bidang bioinformatika dan analisis data genomik. Secara teoritis, penelitian ini diharapkan mampu memperkaya kajian ilmiah terkait penerapan metode seleksi fitur berbasis *fuzzy mutual information* pada data miRNA berdimensi tinggi, khususnya dalam konteks klasifikasi stadium kanker lambung. Secara praktis, penelitian ini diharapkan dapat menjadi referensi dan dasar pengembangan sistem analisis berbasis *machine learning* yang lebih akurat dan adaptif dalam mendukung proses pengambilan keputusan klinis. Dengan demikian, manfaat penelitian ini dirinci sebagai berikut:

1. Memberikan kontribusi ilmiah dalam penerapan teori *fuzzy* dan *mutual information* untuk seleksi fitur pada data biomedis berdimensi tinggi.
2. Menyediakan pendekatan alternatif dalam pemilihan biomarker miRNA yang relevan untuk prediksi stadium kanker lambung berbasis data genomik.
3. Menjadi referensi bagi penelitian selanjutnya yang mengkaji integrasi metode seleksi fitur *fuzzy* dan algoritma *machine learning* pada klasifikasi kanker.

## 1.8 Sistematika Penulisan

Sistematika penulisan laporan ini disusun untuk memberikan gambaran yang terstruktur dan sistematis mengenai alur pembahasan penelitian, sehingga memudahkan pembaca dalam memahami tujuan, metodologi, serta hasil yang diperoleh. Penyusunan sistematika ini diharapkan dapat membantu pembaca mengikuti tahapan penelitian secara runtut, mulai dari latar belakang permasalahan, landasan teori, metode yang digunakan, hingga analisis hasil dan simpulan penelitian. Dengan adanya sistematika penulisan yang jelas, laporan ini tidak hanya berfungsi sebagai dokumentasi ilmiah, tetapi juga sebagai referensi yang informatif bagi peneliti, akademisi, maupun pihak lain yang memiliki ketertarikan pada bidang analisis data genomik dan penerapan *machine learning* dalam kesehatan. Sistematika penulisan laporan adalah sebagai berikut:

- **Bab 1** membahas latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
- **Bab 2** memuat kajian pustaka dan landasan teori yang relevan, meliputi kanker lambung, microRNA, seleksi fitur, teori *fuzzy*, *Fuzzy Mutual Information*, serta algoritma *machine learning*.
- **Bab 3** menjelaskan metodologi penelitian, mencakup sumber data, praproses data, seleksi fitur, penyeimbangan kelas, pembangunan model, skema validasi, dan metrik evaluasi.
- **Bab 4** menyajikan hasil eksperimen, analisis kinerja model, serta pembahasan hasil seleksi fitur dan klasifikasi.
- **Bab 5** berisi kesimpulan penelitian dan saran untuk pengembangan penelitian selanjutnya.