BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi *DNA microarray* dan *high-throughput sequencing* telah menghasilkan data ekspresi gen dalam jumlah yang sangat besar. Kondisi ini mendorong meningkatnya perhatian dalam bidang bioinformatika dan *bioengineering* untuk menganalisis serta mengekstraksi pengetahuan dari data biologis berdimensi tinggi, khususnya dalam konteks penelitian kanker. Salah satu pendekatan yang paling umum digunakan dalam analisis data ekspresi gen adalah metode klasifikasi, yang bertujuan untuk membedakan kondisi biologis tertentu, seperti jaringan normal dan jaringan kanker, berdasarkan pola ekspresi gen.

Namun demikian, data ekspresi gen memiliki karakteristik utama berupa dimensi yang sangat tinggi dengan jumlah sampel yang relatif kecil. Ketidakseimbangan antara jumlah fitur dan jumlah sampel ini menyebabkan banyak metode klasifikasi konvensional menjadi kurang efektif apabila diterapkan secara langsung, karena meningkatnya kompleksitas komputasi serta risiko terjadinya *overfitting*. Oleh karena itu, dalam penelitian data ekspresi gen telah menjadi kesepakatan umum bahwa reduksi dimensi perlu dilakukan sebelum proses klasifikasi untuk meningkatkan stabilitas model dan kualitas prediksi.

Salah satu pendekatan reduksi dimensi yang paling banyak digunakan dalam analisis data ekspresi gen adalah *feature selection*. Berbeda dengan metode ekstraksi fitur, *feature selection* mempertahankan makna biologis dari setiap gen sehingga hasil analisis tetap dapat diinterpretasikan secara biologis. Selain mampu mengurangi kompleksitas komputasi, *feature selection* juga berperan dalam meningkatkan kinerja klasifikasi serta membantu peneliti mengidentifikasi gen-gen yang relevan terhadap mekanisme biologis suatu penyakit.

Secara umum, metode *feature selection* dapat diklasifikasikan ke dalam tiga kategori utama, yaitu *filter*, *wrapper*, dan *embedded methods*. Metode *filter* melakukan seleksi fitur berdasarkan karakteristik statistik data dan hubungan antara fitur dengan kelas target tanpa melibatkan algoritma klasifikasi tertentu. Metode ini relatif efisien secara komputasi dan sesuai untuk data berdimensi tinggi seperti data ekspresi gen. Sebaliknya, *wrapper* dan *embedded methods* mengevaluasi subset fitur berdasarkan performa model klasifikasi, namun umumnya memiliki beban komputasi yang tinggi sehingga kurang sesuai untuk dataset dengan jumlah fitur yang sangat besar [1].

Dalam konteks ini, pendekatan *filter-based feature selection* berbasis teori informasi menjadi menarik karena mampu mengukur relevansi fitur terhadap kelas target secara efisien. Salah satu pendekatan yang dikembangkan adalah *Approximate Conditional Entropy*, yang dirancang untuk mengukur tingkat ketidakpastian informasi suatu fitur dengan mempertimbangkan hubungan kondisional antar fitur. Pendekatan ini memungkinkan pemilihan subset fitur yang informatif tanpa memerlukan proses pelatihan model secara berulang, sehingga lebih sesuai untuk data ekspresi gen berskala besar.

Meskipun berbagai metode seleksi fitur telah diusulkan untuk menangani permasalahan data ekspresi gen berdimensi tinggi, kajian yang secara khusus membahas penggunaan metode seleksi fitur berbasis *approximate conditional entropy* yang dikombinasikan dengan model

klasifikasi *ensemble* modern pada prediksi kanker usus besar masih relatif terbatas. Selain itu, masih diperlukan evaluasi lebih lanjut untuk mengetahui sejauh mana kombinasi tersebut mampu menghasilkan subset gen yang ringkas dengan kinerja klasifikasi yang stabil pada data uji.

Penelitian ini menggunakan pendekatan *Feature Selection using Approximate Conditional Entropy* (FSACE) untuk mereduksi dimensi data ekspresi gen kanker usus besar dengan memilih gen-gen yang memiliki relevansi informasi tinggi terhadap kelas target [1]. Selanjutnya, model klasifikasi *XGBoost* digunakan untuk mengevaluasi kinerja subset fitur terpilih, mengingat kemampuannya dalam menangani data berdimensi tinggi dan kompleks [4]. Pendekatan ini diharapkan dapat memberikan gambaran mengenai potensi kombinasi metode seleksi fitur berbasis teori informasi dan model *ensemble* dalam analisis data ekspresi gen kanker usus besar.

1.2 Rumusan Masalah

Bagian rumusan masalah menjelaskan detail permasalahan utama dalam penelitian ini yang dapat dirumuskan sebagai berikut:

1. Bagaimana penerapan metode *Feature Selection using Approximate Conditional Entropy* (FSACE) dalam menyeleksi fitur ekspresi gen berdimensi tinggi pada data kanker usus besar?
2. Bagaimana kinerja model klasifikasi XGBoost dalam memprediksi kanker usus besar menggunakan fitur hasil seleksi FSACE dibandingkan dengan model tanpa proses seleksi fitur?

1.3 Batasan Permasalahan

Untuk menjaga fokus dan keterukuran penelitian, batasan permasalahan dalam penelitian ini ditetapkan sebagai berikut:

1. Data yang digunakan dalam penelitian ini terbatas pada data ekspresi gen kanker usus besar (*colon cancer*) yang diperoleh dari platform *UCSC Xena Browser* dengan format *gene expression STAR-TPM*.
2. Dataset yang digunakan terdiri dari 515 sampel pasien untuk data ekspresi gen dengan jumlah fitur gen sebanyak 60.660 gen, serta data *phenotype* yang mencakup 562 pasien dengan 86 atribut informasi klinis. Penelitian ini hanya memanfaatkan informasi fenotipe yang relevan dengan proses klasifikasi.
3. Penelitian ini berfokus pada penerapan metode *Feature Selection using Approximate Conditional Entropy* (FSACE) sebagai teknik seleksi fitur, tanpa mengeksplorasi atau membandingkan secara menyeluruh dengan seluruh metode seleksi fitur lain yang tersedia.
4. Model klasifikasi yang digunakan dalam penelitian ini dibatasi pada algoritma *XGBoost* sebagai representasi model pembelajaran mesin berbasis *ensemble*, tanpa melakukan evaluasi pada seluruh jenis model klasifikasi lainnya.
5. Proses evaluasi kinerja model difokuskan pada pengukuran performa klasifikasi berdasarkan metrik evaluasi standar, tanpa mempertimbangkan aspek interpretabilitas biologis lanjutan dari gen-gen terpilih.

6. Penelitian ini tidak membahas analisis lanjutan terkait validasi biologis atau eksperimental terhadap gen hasil seleksi, sehingga hasil penelitian difokuskan pada aspek komputasional dan performa klasifikasi.

1.4 Tujuan Penelitian

Secara khusus, tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan metode *Feature Selection using Approximate Conditional Entropy* (FSACE) untuk melakukan reduksi dimensi fitur pada data ekspresi gen kanker usus besar yang memiliki jumlah fitur sangat besar.
2. Mengevaluasi kinerja model klasifikasi *XGBoost* dalam memprediksi kanker usus besar menggunakan fitur hasil seleksi FSACE berdasarkan metrik evaluasi klasifikasi.

1.5 Urgensi Penelitian

Kompleksitas kanker usus besar adalah salah satu penyakit dengan tingkat mortalitas yang tinggi, di mana perbedaan karakteristik biologis antara stadium awal dan stadium lanjut berperan penting dalam prognosis dan penanganan penyakit. Tingginya dimensi data dan keterbatasan jumlah sampel menjadi tantangan utama dalam proses analisis dan klasifikasi. Tanpa seleksi fitur yang tepat, informasi penting dapat tertutupi oleh *noise* dan meningkatkan risiko *overfitting* pada model prediksi. Oleh karena itu, diperlukan pendekatan komputasional yang mampu mereduksi dimensi data secara efektif sekaligus mempertahankan informasi genetik yang relevan, sehingga analisis ekspresi gen dapat dimanfaatkan secara optimal untuk mendukung prediksi stadium kanker usus besar.

1.6 Luaran Penelitian

Luaran penelitian yang diharapkan adalah sebagai berikut:

1. Laporan ilmiah yang mendokumentasikan metodologi, hasil, dan kesimpulan sebagai referensi penelitian lebih lanjut.
2. Evaluasi model yang mencakup *accuracy, precision, recall, F1-score* dan *confusion matrix* sebagai tolak ukur keberhasilan kombinasi model pelatihan XGBoost dan seleksi fitur menggunakan FSACE.

1.7 Manfaat Penelitian

Berdasarkan tujuan yang telah dirumuskan, penelitian ini diharapkan dapat memberikan manfaat baik secara akademis maupun praktis, khususnya dalam pengembangan metode seleksi fitur dan penerapannya pada klasifikasi data ekspresi gen kanker usus besar. Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan kontribusi akademis berupa evaluasi efektivitas metode *Feature Selection using Approximate Conditional Entropy* (FSACE) dalam mereduksi dimensi data ekspresi gen berdimensi tinggi.
2. Menjadi referensi dalam penerapan metode seleksi fitur berbasis entropi untuk meningkatkan kinerja model klasifikasi XGBoost pada data bioinformatika, khususnya data ekspresi gen kanker usus besar.
3. Menyediakan gambaran empiris mengenai pengaruh seleksi fitur terhadap performa klasifikasi, sehingga dapat digunakan sebagai bahan pertimbangan bagi penelitian selanjutnya dalam bidang klasifikasi data berdimensi tinggi.

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:

1. Bab 1 membahas latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
2. Bab 2 memuat kajian pustaka dan landasan teori yang relevan, meliputi tinjauan teori, kanker usus besar dan klasifikasi kelas, ANOVA, FSACE, dan model XGBoost.
3. Bab 3 menjelaskan metodologi penelitian, mencakup sumber data, praproses data, seleksi fitur, penyeimbangan kelas, dan pembangunan model
4. Bab 4 menyajikan hasil pra-proses data, hasil seleksi fitur, dan pembangunan model klasifikasi
5. Bab 5 berisi kesimpulan penelitian dan saran untuk pengembangan penelitian selanjutnya.