

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Salah satu pemicu utama kematian di dunia yang menjadi permasalahan kesehatan global yang signifikan adalah kanker. Penyakit ini memiliki gejala seperti berkembangnya sel yang tidak normal secara masif dan menyebar ke organ lain dalam tubuh. Peningkatan jumlah kasus kanker setiap tahunnya dipengaruhi oleh berbagai faktor seperti gaya hidup, lingkungan, serta faktor genetik [1].

Di Indonesia, kanker juga menjadi masalah kesehatan masyarakat yang terus meningkat. Salah satu kategori kanker yang paling mematikan adalah kanker paru-paru. Kanker ini sering kali terdeteksi saat sudah mencapai stadium akhir, akibatnya keberhasilan pengobatan menjadi lebih rendah dan berdampak pada rendahnya tingkat kelangsungan hidup pasien [2].

Kanker paru tipe *Lung Adenocarcinoma* (LUAD) merupakan subtype kanker paru yang paling sering terjadi. Tingkat stadium kanker memiliki hubungan yang kuat dengan prognosis pasien, di mana semakin tinggi stadium kanker maka semakin tinggi pula risiko kematian pasien [3]. Namun, pasien dengan stadium yang sama masih dapat menunjukkan luaran *survival* yang berbeda akibat adanya variasi karakteristik molekuler tumor. Kondisi ini menunjukkan bahwa informasi stadium saja belum cukup untuk menggambarkan prognosis setiap pasien secara individual. Oleh karena itu, diperlukan identifikasi biomarker prognostik yang mampu memberikan informasi biologis mengenai risiko *survival* pasien. Data RNA-seq dipilih karena mampu mengukur tingkat ekspresi ribuan gen secara komprehensif sehingga berpotensi mengidentifikasi gen-gen yang berhubungan dengan prognosis pasien. Dengan demikian, identifikasi biomarker prognostik berbasis RNA-seq serta stratifikasi risiko *survival* diharapkan dapat mendukung pengembangan sistem prediksi prognosis yang lebih akurat pada pasien LUAD.

Perkembangan teknologi dalam bidang biologi molekuler memungkinkan analisis kanker melalui pendekatan genomik. Salah satu teknologi yang banyak digunakan adalah RNA *sequencing* (RNA-seq). RNA-seq dapat mengukur ekspresi gen dengan menyeluruh. Data RNA-seq telah terbukti efektif dalam mengidentifikasi biomarker serta membangun model prediksi prognosis pada kanker paru-paru [4].

Saat ini tersedia berbagai dataset genomik kanker berskala besar yang dapat dimanfaatkan untuk penelitian, salah satunya adalah *The Cancer Genome Atlas*. Dataset ini menyediakan data ekspresi gen serta data klinis pasien kanker, termasuk kanker paru-paru tipe LUAD, sehingga memungkinkan analisis dan mencari biomarker prognosis pasien untuk distratifikasi berdasarkan tingkat kelangsung hidup pasien [5].

Seiring dengan pertumbuhan ilmu informatika, pembelajaran mesin (*machine learning*) cukup marak dipakai di bidang kesehatan untuk menganalisis data kompleks. Data RNA-seq memiliki karakteristik berdimensi tinggi dengan jumlah fitur gen yang sangat banyak, sehingga membutuhkan metode yang mampu menangani kompleksitas tersebut. Salah satu algoritma *Machine Learning* yang paling efektif dalam klasifikasi data berdimensi tinggi dan telah banyak digunakan dalam analisis data genomik dan data berdimensi tinggi adalah Support Vector Machine (SVM) [6].

Data RNA-seq memiliki karakteristik *high-dimensional*, di mana jumlah gen yang dianalisis jauh melebihi jumlah sampel, sehingga berpotensi meningkatkan risiko *overfitting*. Oleh karena itu, LASSO digunakan untuk menyeleksi fitur yang paling relevan, sedangkan SVM dipilih sebagai metode klasifikasi karena mampu menangani data berdimensi tinggi secara efektif [7]. Kombinasi kedua metode tersebut diharapkan dapat meningkatkan kemampuan dalam mengidentifikasi biomarker prognostik serta melakukan stratifikasi risiko *survival* pada pasien LUAD. Melalui latar belakang yang sudah didapatkan, penelitian ini memiliki tujuan untuk memanfaatkan data RNA-seq dari dataset TCGA-LUAD untuk melakukan identifikasi biomarker prognostik untuk stratifikasi risiko *survival* pasien kanker paru-paru menggunakan algoritma LASSO dan SVM.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ialah sebagai berikut.

1. Bagaimana proses identifikasi biomarker prognostik berdasarkan data RNA-seq dari dataset TCGA-LUAD dengan menggunakan metode LASSO ?
2. Bagaimana cara membangun model klasifikasi untuk stratifikasi risiko *survival* pasien *lung adenocarcinoma* menggunakan algoritma SVM ?
3. Sejauh mana kinerja model LASSO-SVM dalam mengklasifikasikan risiko *survival* pasien berdasarkan pengukuran evaluasi seperti akurasi, presisi, *recall*, F1-score dan AUC ?

1.3 Batasan Masalah

Penelitian memakai data sekunder dataset RNA-seq TCGA-LUAD (*Lung Adenocarcinoma*) dan tidak melibatkan data dari sumber lain di luar dataset tersebut. Analisis hanya berfokus pada data ekspresi gen serta informasi *survival* pasien, sehingga tidak mencakup integrasi data *multi-omics* seperti proteomik maupun metilasi DNA. Identifikasi biomarker dilakukan menggunakan metode LASSO sebagai teknik seleksi fitur tanpa membandingkannya secara mendalam dengan metode seleksi fitur lainnya. Selanjutnya, model klasifikasi risiko *survival* dikembangkan dengan SVM sebagai pendekatan *prime* tanpa mengeksplorasi algoritma *machine learning* lain secara komprehensif.

Klasifikasi risiko pasien dibatasi menjadi dua kategori, yaitu *low risk* dan *high risk* berdasarkan hasil prediksi model. Evaluasi kapasitas model diukur dengan memakai metrik *accuracy*, *precision*, *recall*, *F1-score*, dan AUC. Penelitian ini juga tidak mencakup validasi eksperimental di laboratorium (*wet lab*), sehingga hasil yang diperoleh sepenuhnya bersifat komputasional berbasis bioinformatika.

1.4 Tujuan Penelitian

Penelitian inipun memiliki tujuan sebagai berikut :

1. Mengidentifikasi biomarker prognostik berbasis data RNA-seq dari dataset TCGA-LUAD menggunakan metode LASSO untuk seleksi fitur gen.
2. Mengembangkan model klasifikasi untuk stratifikasi risiko *survival* pasien *lung adenocarcinoma* (*low risk* dan *high risk*) dengan memanfaatkan SVM
3. Menilai kinerja model LASSO-SVM dalam mengklasifikasikan risiko *survival* pasien berdasarkan berbagai metrik evaluasi.

1.5 Manfaat Penelitian

Penelitian ini juga memiliki manfaat yang dapat dibagikan adalah sebagai berikut :

1. **Akademis**

Menjadi referensi penelitian di bidang bioinformatika dan *machine learning*.

2. **Informatika**

Menunjukkan penerapan *machine learning* pada data genomik berdimensi tinggi.

3. **Kesehatan**

Memberikan gambaran biomarker prognostik untuk stratifikasi risiko hidup pasien kanker paru-paru *Adecarcninoma*.

1.6 Sistematika Penulisan

Sistematika penulisan Skripsi ini adalah sebagai berikut.

Bab 1 PENDAHULUAN

Bab ini berisi latar belakang identifikasi biomarker dan penggunaan LASSO-SVM sebagai algoritmanya dalam stratifikasi risiko *survival* pasien. Selain itu ada batasan masalah serta tujuan dan manfaat dari penelitian ini. Bagian ini menjelaskan arah yang akan dicapai dalam penelitian ini.

Bab 2 LANDASAN TEORI

Bab ini membahas teori-teori yang memiliki kaitan erat dengan penelitian seperti identifikasi biomarker, kanker paru-paru, SVM, RNA-seq dan penelitian lain yang terkait untuk menjadi dasar dalam penelitian.

Bab 3 METODOLOGI PENELITIAN

Bab ini menguraikan metode yang dipakai dalam penelitian seperti pengumpulan data, praproses data, serta pembuatan model dengan LASSO dan implementasi SVM dalam mengidentifikasi biomarker prognostik serta evaluasi model dari hasil yang ada.

Bab 4 HASIL DAN DISKUSI

Bab ini menampilkan hasil dari penelitian berupa gen biomarker yang sudah diidentifikasi serta sudah divalidasi dengan literatur dan data klinis yang sudah ditentukan.

Bab 5 SIMPULAN DAN SARAN

Bab ini menjelaskan kesimpulan dari penelitian yang berupa ringkasan mengenai cara identifikasi biomarker, pembuatan model dan juga hasil evaluasi model yang sudah dibuat hingga saran yang diberikan untuk penelitian selanjutnya.