

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker merupakan salah satu penyebab utama kematian di dunia dan menjadi permasalahan kesehatan global yang signifikan. Berdasarkan laporan dari *World Health Organization*, kanker menyebabkan jutaan kematian setiap tahunnya dan diperkirakan jumlah kasusnya akan terus meningkat seiring dengan pertumbuhan populasi dan perubahan pola hidup masyarakat. Salah satu jenis kanker yang memiliki tingkat mortalitas yang relatif tinggi adalah kanker hati atau *Hepatocellular Carcinoma* (HCC). Penyakit ini dikenal memiliki prognosis yang cukup buruk karena sering kali terdeteksi pada stadium lanjut serta memiliki perkembangan penyakit yang cepat [1, 2].

Secara global, kanker hati termasuk salah satu penyebab utama kematian akibat kanker. Beberapa laporan penelitian menunjukkan bahwa kanker hati menempati peringkat teratas dalam angka kematian terkait kanker di berbagai negara, dengan tingkat kelangsungan hidup lima tahun yang relatif rendah dibandingkan dengan jenis kanker lainnya [3]. Kondisi ini menunjukkan bahwa kanker hati masih menjadi tantangan besar dalam bidang kesehatan, terutama dalam upaya meningkatkan deteksi dini dan menentukan strategi terapi yang tepat bagi pasien [4].

Dalam konteks penanganan kanker, informasi mengenai prognosis atau perkiraan perkembangan penyakit menjadi aspek yang sangat penting. Prognosis berkaitan dengan estimasi kemungkinan kelangsungan hidup pasien setelah diagnosis ditegakkan dan terapi diberikan. Informasi ini dapat membantu tenaga medis dalam menentukan strategi pengobatan yang paling sesuai, mengevaluasi efektivitas terapi, serta memberikan gambaran mengenai harapan hidup pasien. Namun demikian, proses penentuan prognosis pasien kanker masih menghadapi berbagai tantangan karena dipengaruhi oleh banyak faktor, seperti kondisi klinis pasien, stadium penyakit, karakteristik biologis tumor, serta respons terhadap terapi [5]. Hubungan antara berbagai faktor tersebut bersifat kompleks sehingga memerlukan pendekatan analisis yang lebih komprehensif untuk memperoleh hasil prediksi yang lebih akurat [6].

Perkembangan teknologi komputasi dan analisis data telah membuka

peluang baru dalam pemanfaatan data medis untuk mendukung pengambilan keputusan klinis. Salah satu pendekatan yang banyak digunakan adalah teknik *Machine Learning* (ML). Seiring dengan berkembangnya teknologi kecerdasan buatan, sebagian penelitian mulai memanfaatkan pendekatan *Deep Learning* (DL) yang mampu mempelajari pola *non-linear* yang kompleks dalam data berdimensi tinggi. Berbagai literatur melaporkan capaian performa yang impresif dari metode ini; sebagai contoh, model *deep neural network* yang dikembangkan mampu mencapai nilai *concordance index* (*C-index*) di atas 0,70 dalam memodelkan waktu kelangsungan hidup [7]. Selain itu, studi komprehensif lainnya melaporkan bahwa model berbasis DL dapat mencapai akurasi prediksi pada kisaran 75%–85% untuk berbagai jenis kanker [8]. Meskipun DL mampu menghasilkan metrik evaluasi yang tinggi tersebut, pendekatan ini memiliki keterbatasan mendasar dalam konteks medis: kebutuhan akan volume data yang masif, biaya komputasi yang tinggi, dan sifatnya yang *black-box* (kurang interpretabel) [9]. Padahal, dalam praktik klinis, tenaga medis membutuhkan transparansi model untuk memahami variabel biologis apa yang sebenarnya mendorong keputusan prognosis tersebut.

Di sisi lain, metode *machine learning* klasik seperti *Random Forest* (RF) dan *Support Vector Machine* (SVM) masih menjadi andalan dalam analisis bioinformatika. Beberapa penelitian terbaru menunjukkan bahwa metode *machine learning* klasik masih mampu menghasilkan performa prediksi yang sangat kompetitif, bahkan pada beberapa kasus data biologis, akurasinya dapat menyamai model kompleks namun dengan kebutuhan komputasi yang jauh lebih efisien [10, 11]. Keunggulan utama metode-metode ini terletak pada ketahanannya terhadap *overfitting* pada data berdimensi tinggi dan, yang terpenting, tingkat interpretabilitas yang jauh lebih baik. Melalui metode berbasis pohon keputusan seperti RF, peneliti dapat mengekstrak *feature importance* untuk mengidentifikasi *biomarker* genetik spesifik yang paling berpengaruh terhadap kelangsungan hidup pasien [12].

Meskipun demikian, terdapat beberapa tantangan komputasi dalam penerapan ML klasik pada data biomedis. *Pertama*, data genomik seperti ekspresi gen (*RNA-seq*) memiliki dimensi yang sangat tinggi (ribuan gen) namun dengan jumlah sampel pasien yang terbatas (*high-dimensional, low-sample size*), sehingga memerlukan teknik seleksi fitur yang ketat untuk membuang *noise* [13]. *Kedua*, data kelangsungan hidup pasien sering kali mengalami ketidakseimbangan kelas (*class imbalance*). Jumlah pasien dengan risiko kematian tinggi dalam waktu singkat biasanya jauh lebih sedikit dibandingkan pasien yang bertahan hidup [14]. *Ketiga*, prognosis pasien tidak hanya bergantung pada data molekuler, tetapi juga data klinis

(seperti usia, stadium, dan riwayat pengobatan). Penggabungan kedua modalitas data ini (*Early Fusion*) sering kali belum dieksplorasi secara optimal dalam model klasifikasi risiko [15].

Berdasarkan kondisi tersebut, dapat diidentifikasi adanya kesenjangan penelitian (*knowledge gap*) dalam pemanfaatan metode *machine learning* klasik untuk melakukan stratifikasi risiko prognosis *survival* pasien kanker secara biner dengan mempertimbangkan integrasi data multi-modal. Sebagian besar penelitian sebelumnya berfokus pada prediksi *survival* dalam bentuk probabilitas statistik atau menggunakan *deep learning* yang sulit diinterpretasikan, sementara pendekatan klasifikasi risiko biner berbasis *Early Fusion* data genomik dan klinis dengan penanganan *class imbalance* masih perlu dioptimalkan [16, 17]. Pendekatan stratifikasi biner (risiko tinggi vs risiko rendah) dinilai lebih relevan dan mudah diadopsi dalam sistem pendukung keputusan klinis untuk menentukan apakah pasien memerlukan intervensi medis yang agresif atau standar.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan suatu model klasifikasi prognosis *survival* pasien kanker hati dengan memanfaatkan metode *machine learning* klasik (SVM dan Random Forest) melalui pendekatan *Early Fusion* antara data ekspresi gen dan data klinis. Model yang dikembangkan digunakan untuk mengklasifikasikan pasien ke dalam dua kategori risiko berdasarkan ambang batas kelangsungan hidup tiga tahun (1095 hari), yaitu *High Risk* (pasien yang meninggal dalam kurun waktu ≤ 1095 hari) dan *Low Risk* (pasien yang bertahan hidup > 1095 hari). Penelitian ini juga berfokus pada penerapan teknik seleksi fitur untuk mengatasi dimensionalitas data serta penanganan ketidakseimbangan kelas menggunakan metrik *Balanced Accuracy*. Pada akhirnya, penelitian ini tidak hanya bertujuan untuk menghasilkan model prediksi yang *robust*, tetapi juga mengekstrak *biomarker* genetik paling berpengaruh (*Top Feature Importance*) yang dapat diinterpretasikan secara biologis untuk mendukung pengambilan keputusan klinis yang lebih transparan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat dirumuskan beberapa permasalahan sebagai berikut:

1. Bagaimana mengintegrasikan data genomik (*RNA-seq*) dan data klinis melalui pendekatan *Early Fusion*, serta menerapkan seleksi fitur untuk

mengatasi dimensi data yang tinggi dan ketidakseimbangan kelas (*class imbalance*) pada data *survival* pasien kanker hati?

2. Bagaimana kinerja dan optimasi algoritma *Machine Learning* (*Support Vector Machine* dan *Random Forest*) dalam mengklasifikasikan risiko *survival* pasien secara biner dengan mempertimbangkan metrik evaluasi yang sesuai untuk data tidak seimbang?
3. Fitur genomik dan klinis apa saja yang paling berpengaruh (kandidat *biomarker*) dalam memprediksi risiko *survival* berdasarkan tingkat interpretabilitas model yang dibangun?

1.3 Batasan Permasalahan

Agar penelitian ini lebih terarah dan fokus, maka ditetapkan batasan masalah sebagai berikut:

1. Data yang digunakan merupakan data ekspresi gen (*RNA-seq*), data klinis pasien kanker hati (*Hepatocellular Carcinoma*), dan data *survival* atau kelangsungan hidup pasien yang diperoleh dari *database* publik (UCSC Xena / TCGA).
2. Target klasifikasi dibatasi menjadi dua kategori biner berdasarkan ambang batas waktu kelangsungan hidup 3 tahun (1095 hari), yaitu *High Risk* (meninggal ≤ 1095 hari) dan *Low Risk* (bertahan > 1095 hari).
3. Algoritma *Machine Learning* yang digunakan dibatasi pada *Support Vector Machine* (SVM) dan *Random Forest* (RF) dengan penanganan *class imbalance* menggunakan parameter *class weight*.
4. Penggabungan data dilakukan menggunakan pendekatan *Early Fusion* antara fitur genomik yang telah diseleksi (menggunakan *Variance Threshold* dan *SelectKBest*) dan fitur klinis.
5. Evaluasi kinerja model berfokus pada metrik *Balanced Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk mengakomodasi distribusi kelas yang tidak seimbang.
6. Analisis interpretabilitas dan validasi model dibatasi pada ekstraksi *Top 15 Feature Importance* dari model *Random Forest* yang divalidasi melalui

penelusuran literatur biomedis (studi pustaka), tanpa melakukan uji klinis langsung, validasi eksperimental (*wet-lab*), maupun diskusi klinis formal dengan tenaga medis.

7. Hasil penelitian ini murni merupakan pemodelan komputasional dan tidak ditujukan sebagai alat diagnosis atau keputusan klinis final.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan penelitian ini adalah sebagai berikut:

1. Mengintegrasikan data genomik (*RNA-seq*) dan data klinis menggunakan pendekatan *Early Fusion*, serta menerapkan seleksi fitur untuk mengurangi dimensi data yang tinggi dan menangani ketidakseimbangan kelas pada data *survival* pasien kanker hati.
2. Menerapkan dan mengoptimasi algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF), serta mengevaluasi kinerjanya dalam mengklasifikasikan risiko *survival* pasien secara biner menggunakan metrik evaluasi yang sesuai untuk data tidak seimbang.
3. Mengidentifikasi kandidat *biomarker* genomik dan fitur klinis yang paling berpengaruh terhadap prediksi risiko *survival* berdasarkan interpretabilitas model, serta memvalidasinya secara komputasional melalui tinjauan literatur biomedis.

1.5 Manfaat Penelitian

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

1. Manfaat Teoritis

- (a) Memberikan kontribusi pada literatur bioinformatika terkait efektivitas integrasi data multi-modal (*Early Fusion*) dan penanganan *class imbalance* pada data *survival* kanker.
- (b) Menjadi referensi metodologis dalam penerapan *Machine Learning* klasik yang dapat diinterpretasikan (*interpretable ML*) untuk data biomedis berdimensi tinggi dengan jumlah sampel terbatas.

2. Manfaat Praktis

- (a) Menghasilkan daftar kandidat *biomarker* genetik prioritas yang telah ditinjau secara literatur, sehingga dapat mempercepat proses penapisan target terapeutik atau diagnostik bagi peneliti biomedis (riset translasional).
- (b) Memberikan kerangka kerja komputasional (*computational framework*) yang efisien dan transparan bagi praktisi data sains dalam memodelkan prognosis penyakit.
- (c) Memberikan sistem pendukung awal yang dapat mengelompokkan pasien ke dalam stratifikasi risiko (*High/Low Risk*) secara cepat berdasarkan profil molekuler dan klinis mereka.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran umum mengenai isi dari setiap bab yang terdapat dalam laporan skripsi. Adapun sistematika penulisan adalah sebagai berikut:

1. Bab I Pendahuluan

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

2. Bab II Landasan Teori

Bab ini berisi landasan teori yang mendukung penelitian, penelitian terdahulu, serta kerangka pemikiran yang digunakan.

3. Bab III Metodologi Penelitian

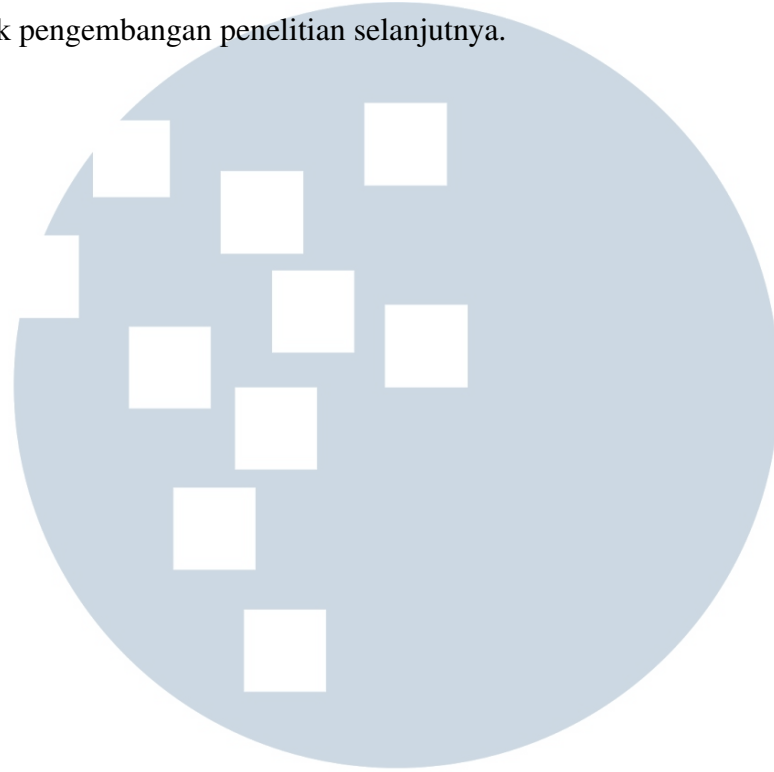
Bab ini menjelaskan tahapan penelitian yang dilakukan, meliputi pengumpulan data, integrasi data genomik dan klinis (*Early Fusion*), praproses data, seleksi fitur, penanganan ketidakseimbangan kelas, penerapan dan optimasi metode klasifikasi, serta evaluasi performa model.

4. Bab IV Hasil dan Pembahasan

Bab ini berisi hasil penerapan metode yang digunakan, analisis performa model, serta interpretasi hasil termasuk identifikasi kandidat biomarker genomik dan validasinya secara komputasional melalui tinjauan literatur biomedis.

5. Bab V Penutup

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



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA