

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker prostat merupakan salah satu jenis kanker yang paling sering dialami oleh pria, terutama pada kelompok usia lanjut [1]. Berdasarkan data Global Cancer Observatory di tahun 2022, kanker prostat menempati peringkat keempat jenis kanker yang paling banyak didiagnosis di dunia, dengan sekitar 1,5 juta kasus. Sementara itu, tingkat kematiannya berada di peringkat kedelapan dengan jumlah kasus hampir mencapai 400.000 [2]. Penelitian sebelumnya mengindikasikan bahwa kejadian kanker prostat bervariasi di antara ras yang berbeda, bahkan setelah mempertimbangkan faktor sosioekonomi, klinis, dan patologis. Khususnya, pria Afrika-Amerika memiliki risiko dua kali lebih tinggi mengalami kanker prostat stadium lanjut dibandingkan pria kulit putih non-Hispanik [3]. Perbedaan ini mengindikasikan adanya faktor genetik yang berperan dalam perkembangan dan agresivitas kanker prostat antar ras.

Tingkat keberhasilan penyembuhan kanker prostat dapat meningkat jika terdeteksi lebih awal [4]. Deteksi umumnya menggunakan biomarker, seperti protein, RNA, DNA, atau metabolit, yang digunakan untuk mengidentifikasi keberadaan penyakit [5]. Biomarker yang umum digunakan adalah Prostate-Specific Antigen (PSA) [6]. PSA mengukur kadar antigen dalam darah, yang cenderung meningkat pada pria dengan kanker prostat. Meskipun PSA berperan penting dalam penentuan stadium dan deteksi dini kanker prostat, penggunaan PSA memiliki keterbatasan, seperti risiko overdiagnosis yang bervariasi antara 1,7% hingga 67% serta tingkat spesifisitas yang rendah [7, 8]. Pengembangan biomarker yang lebih presisi sangat diperlukan untuk mengurangi risiko tersebut.

Kemajuan di bidang teknologi dalam beberapa dekade terakhir telah memungkinkan analisis ekspresi gen dan miRNA menggunakan metode seperti RNA-seq berbasis *Next-Generation Sequencing* (NGS), yang lebih unggul dibandingkan mikroarray atau Sanger [9, 10]. Analisis ini memungkinkan identifikasi pola ekspresi gen yang berhubungan dengan perkembangan kanker prostat [11]. Sementara itu, miRNA mulai dikenal sebagai kandidat biomarker yang menjanjikan karena kestabilannya di cairan tubuh dan kemudahan dalam proses deteksinya [12]. Pada bidang bioinformatika, analisis data ekspresi gen RNA-seq

dan miRNA umumnya diawali dengan pendekatan *Differentially Expressed Genes* (DEG) untuk mengidentifikasi gen atau miRNA yang terekspresi secara signifikan antar kondisi biologis. Tingginya kompleksitas dan jumlah fitur pada data ini mendorong penerapan teknologi komputasi seperti *machine learning*, yang umum digunakan untuk membangun model prediktif melalui proses pelatihan dan validasi [13]. Seleksi fitur digunakan untuk meningkatkan kinerja model dan mempercepat proses komputasi, dengan mengekstraksi pola ekspresi gen dan miRNA yang paling relevan dalam membedakan stadium kanker prostat.

Beberapa penelitian telah menerapkan *machine learning* untuk klasifikasi kanker prostat berbasis RNA-Seq dan miRNA. Singireddy menggunakan 106 sampel RNA-seq dari GEO untuk mengklasifikasikan stadium kanker prostat (T1c–T3b) dengan metode seleksi fitur Minimum Redundancy Maximum Relevance (mRMR) dan model SVM, menghasilkan akurasi antara 80% hingga 99,3% [13]. [14]. Casey memanfaatkan data RNA-Seq dari 546 sampel TCGA untuk mengklasifikasi sel normal atau kanker, menggunakan metode *Mutual Information* (MI), *Principal Component Analysis* (PCA), dan *Recursive Feature Elimination* (RFE) untuk seleksi fitur. Model Logistic Regression kemudian dilatih menggunakan 200 fitur gabungan yang diambil dari data Gene, Exon, Junction, dan Isoform. Model ini memperoleh akurasi 99%, presisi 99,5%, *recall* 100%, dan *F1-score* 99,8% [15]. Kabir menggunakan 550 sampel dari GDC untuk mengklasifikasikan jaringan tumor dan normal. Seleksi fitur dilakukan dengan *Multivariate Correlation*, *Extra-Trees*, dan *Chi-Squared*, lalu dilatih menggunakan model *Random Forest* dengan 36 fitur terpilih. Model tersebut mencapai akurasi, presisi, *recall*, dan *F1-score* masing-masing sebesar 96% [15]. [16]. Singh dkk. menggunakan 106 sampel miRNA dari *miRTarBase* dan *miRBase* untuk membedakan antara *benign prostatic hyperplasia* (BPH) dan kanker prostat. Seleksi fitur dilakukan menggunakan *Mann–Whitney U* (4 miRNA), dan model *Random Forest*, mencapai akurasi 77%, presisi 86%, *recall* 71%, dan *F1-score* 78% [17].

Sebagian besar penelitian masih terbatas pada klasifikasi biner, seperti membedakan individu non-kanker dari penderita kanker prostat, tanpa mempertimbangkan stadium spesifik. Padahal, informasi mengenai stadium memiliki nilai klinis penting karena mencerminkan progresivitas penyakit dan memengaruhi strategi penanganan. Selain itu, variasi genetik antar ras sering diabaikan dalam pengembangan model, meskipun latar belakang genetik yang berbeda dapat menghasilkan profil molekuler yang tidak seragam dan memengaruhi validitas biomarker secara global. Kebanyakan studi juga masih menggunakan

jumlah fitur yang besar, yang berisiko meningkatkan kompleksitas, biaya, dan overfitting, sekaligus mengurangi interpretabilitas. Identifikasi panel biomarker yang lebih ringkas akan lebih bermanfaat secara praktis dalam konteks klinis. Di sisi lain, perbandingan langsung antara efektivitas data ekspresi gen dan miRNA dalam klasifikasi stadium kanker prostat masih jarang dilakukan. Studi komparatif yang lebih mendalam diperlukan untuk menggali potensi masing-masing pendekatan dan mengembangkan strategi integratif yang lebih akurat dan efisien.

Penelitian ini bertujuan mengidentifikasi kelompok biomarker potensial dan mengembangkan model prediktif untuk membedakan kanker prostat stadium awal dan akhir, dengan fokus pada stadium II dan III yang ditemukan pada pria kulit putih non-Hispanik. Pemilihan stadium ini didasarkan pada tingginya prevalensi dalam data serta pentingnya secara klinis sebagai representasi kanker lokal lanjut, serta analisis *Differentially Expressed Genes* (DEG) hanya dapat dilakukan pada dua kelompok. Fokus etnis dipilih untuk meminimalkan bias rasial dalam pemilihan biomarker dan pembangunan model klasifikasi. Untuk mengatasi keterbatasan penelitian sebelumnya, pendekatan ini mengintegrasikan berbagai metode seleksi fitur komprehensif seperti DEG, Minimum Redundancy Maximum Relevance (mRMR), Analysis of Variance (ANOVA), dan Recursive Feature Elimination (RFE) guna memastikan pemilihan fitur yang paling relevan dan non-redundan. Model prediktif kemudian dikembangkan menggunakan algoritma Support Vector Machine (SVM) dan Logistic Regression (LR), yang telah terbukti efektif dalam klasifikasi data biologis. Penelitian ini juga membandingkan performa model berbasis ekspresi gen dan miRNA untuk menentukan pendekatan terbaik dalam klasifikasi stadium kanker prostat. Hasil dari penelitian ini dapat memberikan wawasan awal dalam pemilihan biomarker dan membantu pengembangan metode klasifikasi stadium kanker prostat yang lebih akurat.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang, penelitian ini merumuskan permasalahan sebagai berikut:

1. Bagaimana efektivitas metode seleksi fitur dan model yang digunakan dalam mengklasifikasi stadium kanker prostat pada pria kulit putih non-Hispanik, ditinjau dari metrik evaluasi seperti akurasi, presisi, recall, F1-score makro, dan *ROC-AUC*?

2. Apa saja kelompok biomarker potensial dari fitur ekspresi gen dan miRNA terpilih yang secara kolektif efektif dalam membedakan kanker prostat stadium awal dan lanjut pada pria kulit putih non-Hispanik?
3. Bagaimana perbandingan kinerja model prediktif yang dibangun menggunakan data ekspresi gen *RNA-seq* dan data ekspresi miRNA dalam mengklasifikasikan stadium kanker prostat pada pria kulit putih non-Hispanik, berdasarkan metrik akurasi, presisi, recall, *F1-score* makro, dan *ROC-AUC*?

1.3 Batasan Permasalahan

Dalam penelitian ini, terdapat beberapa pembatasan yang perlu diperhatikan, antara lain sebagai berikut:

- Penelitian ini hanya menggunakan data ekspresi gen dan mirna dari *The Cancer Genome Atlas* (TCGA) yang diakses melalui situs UCSC Xena Browser.
- Data yang digunakan dalam penelitian ini terbatas pada pasien pria kulit putih non-Hispanik yang terdiagnosis kanker prostat.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini dapat dijabarkan sebagai berikut:

- Menyediakan model prediktif berbasis ekspresi gen *RNA-seq* dan miRNA yang dapat memprediksi stadium kanker prostat secara akurat pada pria kulit putih non-Hispanik, dengan evaluasi menggunakan metrik akurasi, presisi, recall, *F1-score* makro, dan *ROC-AUC*.
- Mengidentifikasi kelompok biomarker genetik yang signifikan dari fitur ekspresi gen dan miRNA terpilih yang efektif dalam membedakan kanker prostat stadium awal dan stadium lanjut pada pria kulit putih non-Hispanik.
- Menentukan model prediktif yang memberikan performa lebih baik di antara model berbasis ekspresi gen *RNA-seq* dan ekspresi miRNA, berdasarkan hasil evaluasi dengan metrik akurasi, presisi, recall, *F1-score* makro, dan *ROC-AUC*.

1.5 Manfaat Penelitian

Penelitian ini dapat memberikan beberapa manfaat sebagai berikut:

- Memberikan kontribusi dalam pengembangan model prediktif berbasis ekspresi gen RNA-seq dan miRNA yang dapat digunakan sebagai alat bantu dalam pengambilan keputusan klinis, khususnya bagi pria kulit putih non-Hispanik yang terdiagnosis kanker prostat.
- Menyediakan informasi mengenai kelompok biomarker genetik yang relevan untuk pemantauan perkembangan penyakit dan pengembangan terapi yang lebih personal dalam penanganan kanker prostat.
- Menjadi dasar dalam menentukan jenis data molekuler (ekspresi gen atau miRNA) yang lebih efektif digunakan dalam klasifikasi stadium kanker prostat, sehingga dapat memandu penelitian lanjutan dan aplikasi klinis ke depannya.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran. Sistematika penulisan laporan adalah sebagai berikut:

- **Bab 1 PENDAHULUAN**

Bab ini berisi latar belakang masalah penelitian, rumusan masalah yang ingin dipecahkan, batasan permasalahan agar ruang lingkup penelitian lebih fokus, tujuan penelitian yang ingin dicapai, serta manfaat yang didapatkan dari penelitian ini.

- **Bab 2 LANDASAN TEORI**

Pada bab ini dibahas teori-teori dasar yang mendukung penelitian, termasuk penjelasan tentang kanker prostat, konsep ekspresi gen RNA sequencing dan mikroRNA, serta analisis gen berbeda ekspresi (DEG). Selain itu, bab ini juga mengulas metode seleksi fitur seperti ANOVA dan MRMR, algoritma klasifikasi yang digunakan seperti Support Vector Machine (SVM) dan regresi logistik dengan penalti L2, serta metode evaluasi model.

- **Bab 3 METODOLOGI PENELITIAN**

Bab ini menjelaskan tahapan penelitian secara detail, mulai dari pengambilan dan praproses data RNA-seq dan miRNA, proses seleksi fitur gen dan miRNA, pembuatan model klasifikasi, analisis biomarker potensial yang diperoleh dari fitur terpilih, hingga spesifikasi perangkat keras dan lunak yang digunakan dalam eksperimen.

- **Bab 4 HASIL DAN DISKUSI**

Bab ini memaparkan hasil penelitian, termasuk hasil praproses data, fitur yang terpilih, pembangunan dan evaluasi model klasifikasi, serta analisis biomarker. Selain itu, dilakukan pembahasan terhadap efektivitas metode dan interpretasi hasil yang diperoleh.

- **Bab 5 KESIMPULAN DAN SARAN**

Bab terakhir ini berisi kesimpulan dari seluruh rangkaian penelitian serta saran untuk pengembangan metode atau penelitian selanjutnya agar hasil yang diperoleh dapat diimplementasikan secara lebih luas.

