

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker prostat (prostate adenocarcinoma) merupakan salah satu keganasan paling umum pada pria. Berdasarkan data Global Cancer Statistics (GLOBOCAN) 2020, kanker prostat merupakan kanker kedua paling sering didiagnosis pada pria dengan sekitar 1,41 juta kasus baru dan menjadi penyebab sekitar 375.000 kematian di seluruh dunia pada tahun 2020 [1]. Beban kanker secara global bahkan diproyeksikan meningkat hingga 28,4 juta kasus pada tahun 2040, atau naik 47% dibandingkan tahun 2020 [1]. Tren ini juga konsisten dengan data terkini yang menunjukkan kanker prostat tetap menjadi salah satu penyumbang utama kasus kanker baru pada pria. Di Amerika Serikat, insidensi kanker prostat bahkan meningkat sekitar 3% per tahun pada periode 2014–2019 setelah dua dekade penurunan, setara dengan tambahan sekitar 99.000 kasus baru [2]. Tingginya angka kejadian ini menjadikan deteksi dini dan pemahaman mekanisme molekuler kanker prostat sebagai prioritas dalam penelitian biomedis.

Beban kanker prostat tidak hanya besar secara jumlah, tetapi juga menimbulkan tantangan tersendiri dalam hal pengelolaan. Sebagian besar kasus bersifat indolen dan tumbuh lambat sehingga tidak selalu memerlukan tindakan agresif, sedangkan sebagian kecil bersifat agresif dan berpotensi mematikan. Ketidakmampuan membedakan kedua kelompok ini secara dini kerap berujung pada dua masalah yang berlawanan, yaitu overtreatment pada kasus yang sebenarnya tidak berbahaya, serta keterlambatan penanganan pada kasus yang agresif. Kedua masalah tersebut membawa konsekuensi berupa penurunan kualitas hidup pasien, beban biaya kesehatan yang besar, dan risiko komplikasi dari prosedur yang tidak perlu. Oleh karena itu, kebutuhan akan alat bantu yang mampu memberikan penilaian objektif sejak tahap awal menjadi semakin mendesak, dan pendekatan berbasis data molekuler menawarkan arah yang menjanjikan untuk menjawab kebutuhan tersebut.

Secara klinis, deteksi kanker prostat saat ini sangat bergantung pada pemeriksaan kadar Prostate-Specific Antigen (PSA) dalam darah, pemeriksaan colok dubur (digital rectal examination), pencitraan Multiparametric MRI (mpMRI), dan biopsi jaringan. Meskipun banyak digunakan, pemeriksaan

PSA memiliki keterbatasan berupa spesifisitas yang rendah sehingga sering menimbulkan hasil positif palsu, overdiagnosis, dan overtreatment, karena kadar PSA juga dapat meningkat pada kondisi non-kanker seperti pembesaran prostat jinak dan peradangan [3]. mpMRI meningkatkan akurasi deteksi, tetapi bergantung pada ketersediaan alat, biaya tinggi, serta interpretasi yang bersifat subjektif antar-radiolog [4]. Biopsi jaringan, sebagai standar baku, bersifat invasif dan berisiko menimbulkan komplikasi. Permasalahan ini diperberat oleh kenyataan bahwa sebagian besar kanker prostat bersifat indolen, sedangkan hanya sekitar 25–30% merupakan subtype agresif (skor Gleason 7–10) yang rentan bermetastasis, sehingga dibutuhkan penanda yang mampu membedakan kanker yang signifikan secara klinis dan menghindari prosedur biopsi yang tidak perlu [5]. Keterbatasan-keterbatasan ini mendorong kebutuhan akan penanda molekuler (biomarker) yang lebih objektif dan spesifik [5]. Di sinilah letak urgensi penelitian ini: karena hanya sebagian kecil kanker prostat bersifat agresif namun mematikan, sementara metode konvensional bersifat invasif, mahal, dan kurang spesifik, dibutuhkan penanda molekuler objektif yang dapat mendukung deteksi dini secara lebih tepat dan tidak invasif. Identifikasi gen yang secara konsisten terekspresi berlebih pada jaringan kanker menjadi langkah awal yang penting untuk menyediakan penanda semacam itu sekaligus membuka pemahaman terhadap mekanisme molekuler penyakit [5, 3, 4].

Perkembangan teknologi sekuensing RNA (RNA-Sequencing) memungkinkan pengukuran tingkat ekspresi puluhan ribu gen secara serentak dan memberikan gambaran molekuler kanker yang kaya. Basis data publik seperti The Cancer Genome Atlas untuk kanker prostat (TCGA-PRAD) menyediakan data ekspresi gen beserta metadata klinis yang dapat dimanfaatkan untuk menemukan gen yang terekspresi berlebih (upregulated) pada jaringan kanker [6]. Analisis ekspresi diferensial menggunakan DESeq2, yang memodelkan data hitungan (count) dengan distribusi negative binomial, merupakan metode standar untuk mengidentifikasi gen yang berbeda ekspresinya antara kondisi kanker dan normal [7, 8]. DESeq2 dipilih secara khusus karena dirancang untuk data hitungan RNA-Seq yang bersifat overdispersi (varians melebihi rata-rata) dan menyediakan normalisasi ukuran pustaka serta kontrol false discovery rate melalui koreksi Benjamini-Hochberg, sehingga menjadi rujukan standar pada analisis ekspresi diferensial [7, 8].

Setelah gen penanda diperoleh, model machine learning dapat dilatih untuk mengklasifikasikan jaringan kanker versus normal. Sejumlah penelitian terkini telah menerapkan machine learning pada data ekspresi gen untuk klasifikasi

berbagai jenis kanker, termasuk kanker prostat, dengan hasil yang menjanjikan [9, 10, 11, 12]. Pendekatan tersebut mencakup deep learning untuk diagnosis kanker prostat [13, 14], klasifikasi pan-kanker berbasis ekspresi gen [15], serta klasifikasi data RNA-Seq [16]. Pendekatan ini berpotensi menghasilkan alat bantu diagnosis yang cepat, murah, dan dapat direproduksi. Regresi Logistik dipilih sebagai model utama karena bersifat sederhana, transparan, dan mudah diinterpretasikan sehingga sesuai untuk data berdimensi rendah (tujuh gen), sekaligus menjadi baseline yang kuat untuk diuji terhadap model ensemble yang lebih kompleks [24, 27].

Pemanfaatan data ekspresi gen untuk menemukan penanda kanker memiliki keunggulan dibandingkan pendekatan berbasis citra atau penanda tunggal seperti PSA. Ekspresi gen mencerminkan aktivitas molekuler sel secara langsung, sehingga perubahan yang terjadi pada tahap awal keganasan dapat terdeteksi bahkan sebelum muncul gejala klinis yang nyata. Selain itu, panel gen yang ringkas berpotensi diterjemahkan menjadi uji laboratorium yang lebih murah dan cepat, misalnya melalui reverse transcription quantitative PCR (RT-qPCR), dibandingkan sekuensing menyeluruh. Namun, tantangan utama dari data ekspresi gen adalah dimensinya yang sangat tinggi, yaitu puluhan ribu gen diukur pada jumlah sampel yang jauh lebih sedikit. Kondisi ini, yang dikenal sebagai masalah "kutukan dimensi" (curse of dimensionality), menuntut tahap seleksi fitur yang tepat agar model tidak mempelajari derau dan tetap dapat digeneralisasi. Oleh karena itu, kombinasi antara metode seleksi gen yang berlandaskan statistik yang kuat dan model klasifikasi yang sederhana serta interpretable menjadi pendekatan yang menjanjikan untuk permasalahan ini [9, 6].

Namun, terdapat satu tantangan metodologis yang sering kurang diperhatikan, yaitu kebocoran data (data leakage). Apabila seleksi gen dilakukan menggunakan seluruh sampel, termasuk sampel uji, maka fitur dipilih dengan pengetahuan terhadap label data uji sehingga performa model menjadi terlalu optimis dan tidak mencerminkan kemampuan generalisasi yang sebenarnya [17, 18]. Kebocoran data merupakan salah satu kesalahan umum penerapan machine learning pada bidang genomika yang dapat menyebabkan klaim performa yang menyesatkan [19]. Sebuah studi sistematis bahkan menunjukkan bahwa melakukan seleksi fitur sebelum validasi silang menghasilkan estimasi performa yang bias dibandingkan ketika seleksi dilakukan dengan benar di dalam validasi silang [17]. Oleh karena itu, evaluasi yang ketat dan bebas kebocoran menjadi krusial agar hasil penelitian dapat dipertanggungjawabkan.

Berdasarkan latar belakang tersebut, penelitian ini mengidentifikasi gen

upregulated pada kanker prostat dari data TCGA-PRAD menggunakan DESeq2, kemudian membangun model klasifikasi Regresi Logistik berbasis gen tersebut yang dievaluasi dengan protokol validasi silang bebas kebocoran (seleksi gen di dalam lipatan pelatihan, SMOTE per-lipatan, dan pembagian per-pasien). Kontribusi penelitian ini adalah menyajikan tanda tangan gen yang stabil dan tervalidasi secara biologis serta estimasi performa yang jujur dan dapat direplikasi.

Perlu ditegaskan perbedaan antara identifikasi gen upregulated dan diagnosis kanker agar tujuan penelitian ini jelas. Identifikasi gen upregulated bertujuan menemukan gen penanda yang ekspresinya secara konsisten meningkat pada jaringan kanker dibandingkan jaringan normal, yakni pada tataran penemuan kandidat biomarker molekuler. Sementara itu, diagnosis kanker merupakan keputusan klinis menyeluruh yang melibatkan pemeriksaan patologi, pencitraan, riwayat pasien, dan penilaian dokter. Dengan demikian, tujuan penelitian ini adalah mengidentifikasi panel gen penanda serta membuktikan secara komputasional (proof of concept) bahwa panel tersebut mampu membedakan jaringan kanker dan normal, bukan menghasilkan alat diagnosis mandiri yang menggantikan pemeriksaan klinis [5].

1.2 Rumusan Masalah

1. Gen apa saja yang terekspresi berlebih (upregulated) secara signifikan pada jaringan kanker prostat dibandingkan jaringan normal berdasarkan analisis DESeq2 pada data TCGA-PRAD?
2. Bagaimana performa model klasifikasi Regresi Logistik berbasis gen upregulated tersebut, diukur dengan ROC AUC, F1-score, dan recall menggunakan validasi silang bebas kebocoran, serta bagaimana perbandingannya terhadap model stacking ensemble yang lebih kompleks?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Mengidentifikasi gen upregulated pada kanker prostat menggunakan analisis DESeq2 pada data TCGA-PRAD.
2. Mengukur performa model klasifikasi Regresi Logistik berbasis gen upregulated menggunakan ROC AUC, F1-score, dan recall dengan validasi

silang bebas kebocoran, serta membandingkannya dengan model stacking ensemble.

1.4 Batasan Masalah

1. Data yang digunakan adalah TCGA-PRAD kohort ras kulit putih (white) sebanyak 458 sampel.
2. Nilai ekspresi htseq_counts berformat $\log_2(\text{count}+1)$ ditransformasi balik menjadi raw counts sebelum analisis DESeq2.
3. Klasifikasi bersifat biner, yaitu jaringan kanker versus jaringan normal.
4. Kriteria seleksi gen upregulated adalah $\text{padj} < 0,05$, $\text{baseMean} \geq 10$, dan $\log_2\text{FC} > 0,35$.
5. Validasi belum dilakukan pada dataset eksternal independen.

Batasan-batasan tersebut ditetapkan secara sadar agar penelitian tetap terfokus dan dapat diselesaikan dengan sumber daya yang tersedia, sekaligus menjaga validitas internal hasil. Pembatasan pada satu kohort populasi dan satu sumber data memungkinkan kontrol yang lebih baik terhadap faktor perancu, meskipun menukar sebagian generalisasi. Fokus pada klasifikasi biner antara jaringan kanker dan normal juga dipilih agar permasalahan terdefinisi dengan jelas dan hasilnya mudah diinterpretasikan, sebelum penelitian dapat diperluas ke persoalan yang lebih kompleks seperti penilaian tingkat keganasan. Dengan menyatakan batasan secara eksplisit, ruang lingkup dan klaim penelitian menjadi jelas, dan pembaca dapat menilai sejauh mana hasil ini berlaku serta bagian mana yang masih memerlukan penelitian lanjutan.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah sebagai berikut.

1. Manfaat teoritis: memberikan daftar gen penanda kanker prostat yang stabil secara statistik dan tervalidasi secara biologis sebagai dasar penelitian lanjutan.

2. Manfaat metodologis: menunjukkan penerapan protokol validasi silang bebas kebocoran sebagai praktik baik dalam analisis machine learning pada data ekspresi gen.
3. Manfaat praktis: menyediakan model klasifikasi sederhana namun akurat yang berpotensi mendukung pengembangan alat bantu diagnosis kanker prostat.

Ketiga manfaat tersebut saling berkaitan dan mendukung tujuan jangka panjang penelitian biomarker. Secara teoritis, daftar gen penanda yang stabil dapat menjadi titik awal bagi penelitian biologi molekuler untuk menelusuri peran gen-gen tersebut dalam patogenesis kanker prostat. Secara metodologis, protokol evaluasi bebas kebocoran yang ditunjukkan pada penelitian ini dapat diadopsi oleh peneliti lain sebagai kerangka kerja untuk memastikan hasil yang dilaporkan tidak terlalu optimis, sehingga meningkatkan kualitas dan keterpercayaan penelitian machine learning di bidang genomika secara umum. Secara praktis, model dan panel gen yang dihasilkan dapat menjadi landasan bagi pengembangan uji diagnostik yang lebih terjangkau, yang pada akhirnya diharapkan dapat membantu proses penapisan dan pengambilan keputusan klinis. Dengan demikian, penelitian ini tidak hanya berkontribusi pada aspek teknis pemodelan, tetapi juga pada praktik penelitian yang lebih baik dan potensi penerapan yang bermanfaat.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun sebagai berikut.

1. **BAB 1 PENDAHULUAN.** Menjelaskan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.
2. **BAB 2 LANDASAN TEORI.** Membahas penelitian terdahulu serta teori yang mendasari penelitian, meliputi kanker prostat, data ekspresi gen RNA-Seq, analisis ekspresi diferensial DESeq2, metode klasifikasi, validasi silang, kebocoran data, dan metrik evaluasi.
3. **BAB 3 METODOLOGI PENELITIAN.** Menguraikan tahapan penelitian mulai dari pengumpulan data, praproses, seleksi gen, penanganan ketidakseimbangan kelas, pemodelan, hingga evaluasi bebas kebocoran, beserta rincian implementasi yang dapat direplikasi.

4. **BAB 4 HASIL DAN DISKUSI.** Menyajikan hasil identifikasi gen upregulated, stabilitas seleksi gen, performa model klasifikasi, perbandingan model, serta diskusi mengenai validitas dan keterbatasan penelitian.
5. **BAB 5 SIMPULAN DAN SARAN.** Berisi simpulan yang menjawab tujuan penelitian dan saran untuk penelitian selanjutnya.

