

BAB 2

LANDASAN TEORI

2.1 Kanker

Kanker adalah kondisi dimana perkembangan sel tidak normal terjadi secara tidak terkendali yang menyerang jaringan di sekitar sel tersebut dan menjalar di bagian lain dari tubuh manusia karena adanya proses metastasis. Penyakit ini sudah menjadi satu dari banyaknya penyebab utama kematian di dunia dan terus mengalami peningkatan setiap tahunnya [9]. Menurut laporan global kesehatan, peningkatan kasus kanker dipengaruhi banyak faktor seperti pola hidup buruk, faktor lingkungan, paparan zat karsinogen, konsumsi alkohol, dan kebiasaan merokok [10].

Di Indonesia, salah satu isu kesehatan di masyarakat yang cukup berat adalah kanker. Tingginya angka kematian akibat kanker dipengaruhi oleh keterlambatan diagnosis dan rendahnya kesadaran masyarakat terhadap deteksi dini penyakit kanker [11].

2.2 Kanker Paru-Paru

Kanker paru adalah jenis kanker dengan tingkat kematian tertinggi di dunia. Penyakit ini muncul karena adanya pertumbuhan sel abnormal di jaringan paru-paru yang berkembang secara tidak terbenyung [12]. Faktor yang menyebabkan kanker paru-paru adalah kebiasaan merokok, akan tetapi, faktor lain seperti polusi udara, faktor genetik, dan paparan bahan kimia tertentu juga memperparah risiko terjangkitnya manusia dengan penyakit ini [13].

Secara umum, kanker paru-paru dibagi menjadi dua kelompok utama *Non-Small Cell Lung Cancer* (NSCLC) dan *Small Cell Lung Cancer* (SCLC). NSCLC adalah subtype kanker paru-paru yang paling sering ditemukan dan mencakup beberapa subtype, salah satunya adalah *Lung Adenocarcinoma* (LUAD) [14]. Kanker paru-paru dibagi menjadi beberapa stadium berdasarkan tingkat penyebaran penyakit. Stadium I menunjukkan kanker masih berada pada paru-paru, sedangkan stadium IV menunjukkan kanker sudah menjangkau bagian-bagian lain dalam tubuh [15]. Perbedaan stadium tersebut tidak hanya menggambarkan tingkat

keparahan penyakit, tetapi juga memiliki keterkaitan dengan perubahan ekspresi biomarker prognostik yang dapat memengaruhi luaran klinis pasien, termasuk tingkat *survival*. Seiring bertambahnya stadium kanker, perubahan molekuler serta pola ekspresi gen umumnya menjadi semakin kompleks, yang turut berperan sebagai biomarker dalam menentukan prognosis pasien.

2.2.1 Lung Adenocarcinoma (LUAD)

Lung Adenocarcinoma (LUAD) merupakan salah satu sub tipe NSCLC yang berasal dari sel epitel penghasil lendir pada paru-paru. LUAD adalah sub tipe kanker paru yang banyak ditemukan pada pasien non-perokok maupun perokok. Tingkat keberhasilan terapi sangat dipengaruhi oleh stadium penyakit saat diagnosis dilakukan. Karena tingkat heterogenitas biologis yang tinggi, analisis data molekuler seperti ekspresi gen menjadi penting untuk membantu klasifikasi stadium dan prediksi prognosis pasien.

LUAD umumnya berkembang pada bagian perifer paru-paru dan kebanyakan tidak muncul gejala yang berbahaya pada tahap awal. Akibatnya, banyak kasus baru terdiagnosis ketika penyakit telah memasuki stadium lanjut sehingga peluang keberhasilan terapi menjadi lebih rendah [16]. Gejala yang dapat muncul pada penderita LUAD meliputi batuk yang tak kunjung sembuh, nyeri pada dada, sulit dalam bernapas, berat badan yang menurun, dan batuk berdarah [17].

Karakteristik biologis LUAD sangat kompleks karena melibatkan berbagai perubahan genetik dan molekuler yang memengaruhi pertumbuhan, proliferasi, dan penyebaran sel kanker. Beberapa gen yang diketahui sering mengalami mutasi pada pasien LUAD antara lain *EGFR*, *KRAS*, *TP53*, *ALK*, dan *BRAF* [18]. Mutasi pada gen-gen tersebut dapat memengaruhi perkembangan penyakit serta respons pasien terhadap terapi tertentu.

Selain memiliki variasi genetik yang tinggi, LUAD juga menunjukkan tingkat heterogenitas tumor yang signifikan. Heterogenitas tumor mengacu pada adanya perbedaan karakteristik biologis antar sel kanker dalam satu pasien maupun antar pasien yang berbeda [19]. Kondisi ini menyebabkan setiap pasien dapat memiliki perkembangan penyakit, respons terapi, dan prognosis yang berbeda meskipun berada pada stadium klinis yang sama.

Berdasarkan tingkat penyebarannya, LUAD diklasifikasikan ke dalam beberapa stadium, mulai dari stadium I hingga stadium IV. Stadium I dan II umumnya menunjukkan bahwa kanker masih terbatas pada paru-paru, sedangkan stadium III dan IV menunjukkan kanker sudah tersebar pada jaringan sekitar atau organ lain di luar paru-paru [20]. Penentuan stadium memiliki peran penting dalam pemilihan metode pengobatan dan evaluasi prognosis pasien.

Tingginya kompleksitas biologis serta heterogenitas pada LUAD mendorong berkembangnya berbagai penelitian yang memanfaatkan data molekuler untuk memahami mekanisme penyakit secara lebih mendalam. Analisis terhadap profil ekspresi gen dan biomarker molekuler merupakan salah satu pendekatan yang sering dipakai untuk identifikasi karakteristik tumor, klasifikasi stadium, serta prediksi tingkat kelangsungan hidup pasien [21].

2.3 RNA-seq

RNA Sequencing (RNA-seq) merupakan teknologi pemetaan genom moderen yang dipakai untuk menguji tingkat ekspresi gen pada suatu spesimen [22]. Data RNA-seq memungkinkan peneliti mengidentifikasi:

- Tingkat ekspresi gen.
- Gen yang mengalami diferensial ekspresi.
- Biomarker penyakit.
- Mekanisme molekuler kanker.

Selain itu RNA-seq memiliki karakteristik yaitu:

- Berdimensi tinggi (*high dimensional data*).
- Memiliki jumlah fitur yang sangat besar.
- Rentan terhadap *noise*.
- Membutuhkan proses normalisasi sebelum analisis.

Selain informasi mengenai tingkat ekspresi gen, data RNA-seq juga menyediakan anotasi biologis, seperti *Gene Symbol*, *Gene Name*, dan *Gene Type*. Anotasi tersebut dapat diperoleh melalui basis data *MyGene.info*, yang memungkinkan konversi *Ensembl Gene ID* menjadi identitas gen yang lebih mudah dikenali dan diinterpretasikan [23]. *Gene Type* merupakan klasifikasi gen

berdasarkan karakteristik dan fungsi biologisnya. Beberapa tipe gen yang umum ditemukan pada data RNA-seq meliputi *protein-coding*, yaitu gen yang menghasilkan protein melalui proses transkripsi dan translasi, serta *small nuclear RNA (snRNA)*, yaitu RNA non-koding yang berperan dalam proses *RNA splicing*. Dalam penelitian ini, biomarker yang berhasil diidentifikasi sebagian besar merupakan gen bertipe *protein-coding*, sedangkan satu biomarker termasuk ke dalam tipe *snRNA*. Jumlah fitur yang sangat besar pada data RNA-seq menyebabkan tidak semua gen memiliki kontribusi yang signifikan terhadap proses klasifikasi. Oleh karena itu, dilakukan *variance filtering* sebagai tahap awal seleksi fitur untuk menyaring gen berdasarkan tingkat variasi ekspresinya antar sampel. Tahap ini bertujuan menghilangkan gen dengan variansi rendah yang cenderung kurang informatif serta mempertahankan gen dengan variansi tinggi yang memiliki potensi lebih besar dalam membedakan karakteristik setiap sampel [24]. Nilai variansi setiap gen dihitung menggunakan rumus 2.1 berikut.

$$Var(X) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2.1)$$

dengan:

x_i = nilai ekspresi gen pada sampel ke- i

$\bar{x}$ = rata-rata nilai ekspresi gen

n = jumlah sampel

Nilai variansi yang tinggi menunjukkan bahwa suatu gen memiliki tingkat variasi ekspresi yang besar antar sampel sehingga dianggap lebih informatif dalam proses analisis. Sebaliknya, gen dengan nilai variansi rendah cenderung memiliki pola ekspresi yang relatif seragam sehingga dieliminasi untuk mengurangi kompleksitas data serta meningkatkan efisiensi komputasi sebelum dilakukan proses seleksi fitur menggunakan LASSO [24].

2.4 Biomarker

Biomarker adalah karakteristik biologis yang dapat diukur secara objektif dan dimanfaatkan sebagai indikator untuk menggambarkan kondisi fisiologis, proses patologis, maupun respons individu terhadap suatu intervensi atau pengobatan [25]. Dalam bidang onkologi, biomarker memiliki peran penting dalam mendukung proses deteksi penyakit, penentuan prognosis, serta pemilihan terapi yang tepat sehingga dapat menunjang penerapan pengobatan yang lebih terarah [26].

Berdasarkan fungsi klinisnya, biomarker dapat dibedakan menjadi biomarker diagnostik, biomarker prognostik, dan biomarker prediktif. Biomarker diagnostik dimanfaatkan untuk mendeteksi, mengidentifikasi, atau mengonfirmasi keberadaan suatu penyakit sehingga dapat membantu proses penegakan diagnosis secara lebih akurat. Pada kanker paru, contoh biomarker diagnostik yang banyak digunakan meliputi mutasi *epidermal growth factor receptor (EGFR)* dan *anaplastic lymphoma kinase (ALK rearrangement)* untuk mengidentifikasi subtype molekuler *non-small cell lung cancer (NSCLC)*, serta *carcinoembryonic antigen (CEA)* yang berperan sebagai penanda tambahan dalam evaluasi penyakit [27].

Biomarker prognostik digunakan untuk memperkirakan luaran klinis pasien, seperti risiko perkembangan penyakit, peluang terjadinya kekambuhan, maupun tingkat *overall survival*, tanpa mempertimbangkan jenis terapi yang diterima. Biomarker ini merefleksikan karakteristik biologis tumor sehingga dapat dimanfaatkan untuk mengelompokkan pasien berdasarkan tingkat risiko dan membantu penentuan prognosis penyakit. Pada *Lung Adenocarcinoma (LUAD)*, beberapa biomarker prognostik yang telah banyak dilaporkan antara lain *DKK1*, *VEGFA*, dan *KRAS*, yang diketahui berkaitan dengan agresivitas tumor serta luaran klinis pasien [27].

Adapun biomarker prediktif berfungsi untuk memperkirakan kemungkinan respons pasien terhadap suatu terapi tertentu sehingga dapat mendukung pemilihan strategi pengobatan yang lebih efektif. Biomarker jenis ini menjadi salah satu komponen penting dalam penerapan *precision medicine* karena memungkinkan terapi dipilih sesuai dengan karakteristik molekuler masing-masing pasien. Contoh biomarker prediktif pada kanker paru meliputi mutasi *EGFR* sebagai indikator

respons terhadap terapi *EGFR tyrosine kinase inhibitor (EGFR-TKI)*, ekspresi *programmed death-ligand 1 (PD-L1)* sebagai indikator respons terhadap imunoterapi, serta *ALK rearrangement* yang digunakan sebagai dasar pemberian terapi *ALK inhibitor* [27].

Perkembangan metode *bioinformatics* dan *machine learning* telah memungkinkan proses identifikasi biomarker dilakukan secara komputasional. Pendekatan ini sangat efektif untuk mengurus data yang mempunyai dimensi tinggi seperti *RNA-seq*, yang jumlah gennya jauh lebih banyak dibandingkan jumlah sampel. Oleh karena itu, diperlukan metode seleksi fitur untuk mengekstraksi informasi gen yang paling relevan terhadap outcome klinis [28].

Salah satu metode yang sering dipakai adalah LASSO (*Least Absolute Shrinkage and Selection Operator*), yang bekerja dengan memberikan penalti pada koefisien variabel sehingga hanya fitur yang paling signifikan yang dipertahankan. Selanjutnya, model klasifikasi semacam SVM dimanfaatkan untuk memisahkan pasien berdasarkan risiko *survival*, seperti kelompok *low risk* dan *high risk*, berdasarkan pola ekspresi gen yang telah diseleksi [29]. Dengan demikian, biomarker prognostik tidak hanya sebagai penanda biologis saja, tetapi juga menjadi dasar penting dalam pengembangan model prediksi berbasis *machine learning* untuk stratifikasi risiko pasien LUAD secara lebih akurat dan objektif [30].

Dalam penelitian biomarker prognostik kanker paru, penggunaan karakteristik klinis sebagai variabel pendukung diperlukan untuk mengevaluasi kesesuaian antara biomarker yang teridentifikasi dengan kondisi klinis pasien. Variabel klinis yang dipilih pada penelitian ini meliputi stadium patologis (*AJCC Pathologic Stage*), usia saat diagnosis (*Age at Diagnosis*), dan riwayat merokok yang direpresentasikan oleh *Pack Years Smoked*. Ketiga variabel tersebut dipilih karena telah banyak dilaporkan memiliki hubungan dengan perkembangan penyakit dan tingkat kelangsungan hidup pasien *Lung Adenocarcinoma (LUAD)*.

2.5 Stadium Patologis

Stadium patologis merupakan salah satu indikator utama yang digunakan untuk menggambarkan tingkat perkembangan kanker pada pasien. Sistem *American Joint Committee on Cancer (AJCC)* mengelompokkan kanker berdasarkan ukuran tumor primer, keterlibatan kelenjar getah bening regional, serta keberadaan metastasis. Informasi stadium memiliki peranan penting dalam menentukan prognosis dan strategi penatalaksanaan pasien [31].

Sejumlah penelitian menunjukkan bahwa pasien dengan stadium lanjut cenderung memiliki peluang kelangsungan hidup yang lebih rendah dibandingkan pasien pada stadium awal. Hal tersebut disebabkan oleh meningkatnya penyebaran sel kanker dan kompleksitas terapi yang diperlukan pada stadium lanjut [32]. Oleh karena itu, hubungan antara ekspresi biomarker dengan stadium patologis dapat digunakan untuk menilai apakah biomarker yang teridentifikasi berkaitan dengan progresivitas penyakit.

2.5.1 Usia saat Diagnosis

Usia saat diagnosis merupakan faktor klinis yang sering dikaitkan dengan prognosis pasien kanker paru. Peningkatan usia umumnya berhubungan dengan menurunnya fungsi fisiologis tubuh, meningkatnya jumlah penyakit penyerta, serta berkurangnya toleransi terhadap terapi kanker [33].

Beberapa studi melaporkan bahwa pasien dengan usia yang lebih tua memiliki tingkat mortalitas yang lebih tinggi dibandingkan kelompok usia yang lebih muda. Selain itu, perubahan biologis yang terjadi seiring bertambahnya usia juga dapat memengaruhi ekspresi gen serta respons imun terhadap perkembangan tumor [34]. Oleh karena itu, analisis hubungan biomarker dengan usia diharapkan dapat memberikan gambaran mengenai keterkaitan biomarker terhadap faktor prognostik yang telah dikenal secara klinis.

2.5.2 Riwayat Merokok

Merokok merupakan faktor risiko utama terjadinya kanker paru dan berkontribusi terhadap sebagian besar kasus *Lung Adenocarcinoma*. Paparan zat karsinogen dalam asap rokok dapat menyebabkan akumulasi mutasi genetik, kerusakan DNA, serta perubahan pola ekspresi gen yang mendukung proses karsinogenesis [35].

Pada penelitian ini, riwayat merokok direpresentasikan menggunakan variabel *Pack Years Smoked*, yaitu ukuran paparan kumulatif rokok yang dihitung berdasarkan jumlah bungkus rokok yang dikonsumsi per hari dikalikan dengan lama waktu merokok dalam tahun. Variabel ini dianggap lebih representatif dibandingkan hanya menggunakan status perokok atau bukan perokok karena mampu menggambarkan tingkat paparan rokok secara kuantitatif [36].

Berbagai penelitian menunjukkan bahwa paparan rokok yang lebih tinggi berkaitan dengan peningkatan beban mutasi tumor (*tumor mutational burden*), perubahan profil ekspresi gen, serta prognosis yang lebih buruk pada pasien kanker paru [37]. Oleh sebab itu, evaluasi hubungan antara biomarker yang teridentifikasi dengan *Pack Years Smoked* dapat digunakan untuk memperkuat validasi biologis biomarker prognostik yang diperoleh.

2.6 Median Survival Time pada Kanker Paru-Paru

Median survival time merupakan ukuran statistik yang digunakan dalam studi kanker untuk menggambarkan titik waktu ketika separuh pasien telah mengalami kejadian (*event*) seperti kematian, sedangkan separuh lainnya masih bertahan hidup. Pada kasus kanker paru-paru, khususnya *lung adenocarcinoma*, ukuran ini sering dijadikan indikator prognosis karena lebih robust dibandingkan nilai rerata, terutama pada data *survival* yang bersifat tidak simetris serta mengandung data tersensor (*censored data*) [38].

Nilai *median survival time* umumnya dihitung melalui pendekatan *survival analysis*, yaitu metode statistik yang dipakai dalam menganalisis lama waktu hingga kejadian tertentu terjadi, seperti kematian atau kekambuhan penyakit. Dalam penelitian kanker, metode ini banyak digunakan karena mampu menggambarkan pola ketahanan hidup pasien berdasarkan periode observasi secara lebih informatif [38].

2.7 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) merupakan proyek penelitian kanker berskala besar yang berfokus pada pengumpulan, pengolahan, dan analisis data genomik dari berbagai jenis kanker manusia [40]. TCGA menjadi salah satu sumber data utama dalam bidang bioinformatika dan *cancer genomics* karena menyediakan data yang terintegrasi secara *multi-platform*, sehingga memungkinkan analisis hubungan antara perubahan molekuler dan kondisi klinis pasien secara lebih komprehensif [41].

Di dalam TCGA, terdapat beberapa jenis data biologis utama. Pertama adalah data ekspresi gen (*gene expression data*) yang menunjukkan tingkat aktivitas gen pada sampel tumor. Adanya teknologi RNA-seq membuat pengukuran ekspresi gen secara bersamaan membuat data ini bisa didapatkan, sehingga sangat berguna untuk mengidentifikasi gen yang berpotensi sebagai biomarker [42].

Kedua adalah data mutasi (*mutation data*), yaitu informasi mengenai perubahan urutan DNA yang terjadi pada sel kanker. Perubahan ini dapat

memengaruhi peran gen tertentu, salah satunya adalah gen yang berperan dalam pengaturan pertumbuhan sel, sehingga berkontribusi terhadap perkembangan kanker.

Ketiga adalah data klinis (*clinical data*) yang mencakup berbagai informasi pasien seperti usia, jenis kelamin, stadium kanker, jenis terapi, serta data *survival*. Data ini digunakan untuk menghubungkan kondisi molekuler dengan luaran klinis pasien, termasuk tingkat kelangsungan hidup (*survival rate*) dan prognosis.

Keempat adalah data metilasi DNA (*DNA methylation data*), yaitu perubahan epigenetik yang dapat memengaruhi ekspresi gen tanpa mengubah urutan DNA. Perubahan pola metilasi ini sering ditemukan pada kanker dan berperan dalam regulasi aktivasi maupun penekanan gen tertentu [43].

2.7.1 TCGA-LUAD

TCGA-LUAD merupakan salah satu sub-dataset dari *The Cancer Genome Atlas* yang berfokus pada kanker *lung adenocarcinoma* (LUAD). Dataset ini banyak dimanfaatkan dalam penelitian bioinformatika dan *machine learning* karena menyediakan data yang bersifat multidimensi, sehingga memungkinkan analisis yang lebih komprehensif antara karakteristik molekuler dan kondisi klinis pasien [44].

Salah satu komponen utama dalam *TCGA-LUAD* adalah data *RNA sequencing* yang merepresentasikan ekspresi gen dalam sampel jaringan tumor pasien. Data ini memiliki peran penting dalam studi *omics* karena dapat digunakan untuk mengidentifikasi gen yang berpotensi sebagai biomarker prognostik maupun diagnostik [45].

Selain itu, *TCGA-LUAD* juga menyediakan informasi stadium kanker yang menunjukkan tingkat perkembangan penyakit pada pasien, mulai dari stadium awal hingga stadium lanjut. Informasi ini membantu dalam mengevaluasi keterkaitan antara progresivitas kanker dengan perubahan ekspresi gen serta *outcome* klinis pasien [46].

Dataset ini juga mencakup data *survival* pasien, seperti *overall survival (OS)*, yang digunakan untuk menilai ketahanan hidup pasien dalam rentang waktu tertentu. Data ini menjadi komponen penting dalam analisis *survival analysis* serta proses stratifikasi risiko pasien pada penelitian kanker [47].

Di samping itu, *TCGA-LUAD* juga menyediakan berbagai data klinis lainnya, seperti usia, jenis kelamin, riwayat merokok, serta informasi terapi yang diberikan. Variabel-variabel klinis ini sering dimanfaatkan sebagai faktor pendukung dalam analisis *machine learning* guna meningkatkan performa model prediksi prognosis pasien [48].

2.8 Machine Learning

Machine Learning adalah cabang dari *Artificial Intelligence* yang membuat komputer dapat mengkaji pola yang terbentuk dari data dan memprediksi tanpa diatur secara spesifik [49]. Berdasarkan cara belajarnya, *machine learning* dibagi dalam beberapa jenis yaitu *Supervised*, *Unsupervised* dan *Reinforcement Learning*.

2.8.1 Supervised Learning

Supervised learning adalah salah satu cara dalam *machine learning* yang memakai data yang sudah berlabel (*labeled data*) pada proses pelatihannya. Data berlabel adalah data yang telah memiliki informasi mengenai hasil atau kelas yang ingin diprediksi. Tujuan utama dari *supervised learning* adalah membangun model yang mampu mempelajari keterkaitan antar variabel *input* dan variabel keluaran (*output*) yang mampu memprediksi data baru yang belum pernah dilihat sebelumnya.

Pada proses pembelajaran, algoritma akan memproses pola-pola yang terdapat di data latih (*training data*) untuk membuat suatu fungsi prediksi. Setelah model selesai dilatih, model dapat dimanfaatkan untuk mengklasifikasikan atau memprediksi nilai dari data baru. Metode *supervised learning* umumnya dibagi menjadi dua kategori, yaitu klasifikasi dan regresi. Klasifikasi dipakai untuk memprediksi data masuk pada kategori tertentu, sedangkan regresi dimanfaatkan untuk menghitung data numerik yang bersifat kontinu [50].

Dalam bidang kesehatan, khususnya pada penelitian kanker, *supervised learning* sering digunakan untuk mengklasifikasikan jenis kanker, menentukan stadium penyakit, memprediksi prognosis pasien, serta mengidentifikasi risiko penyakit berdasarkan data klinis maupun data molekuler. Beberapa algoritma yang termasuk dalam kategori *supervised learning* adalah SVM, *Decision Tree*, *Random Forest*, *Naïve Bayes*, dan *Artificial Neural Network* [51].

2.8.2 Unsupervised Learning

Unsupervised learning adalah cara *machine learning* yang bekerja dengan data tanpa label (*unlabeled data*). Berbeda dengan *supervised learning*, pada metode ini algoritma tidak diberikan informasi mengenai kelas atau hasil yang diharapkan. Fungsi *unsupervised learning* adalah mencari pola, hubungan atau kelompok yang tidak terlihat yang terdapat pada data secara otomatis [52].

Metode ini banyak digunakan untuk eksplorasi data (*data exploration*), pengelompokan data (*clustering*), reduksi dimensi (*dimensionality reduction*), serta analisis hubungan antar variabel. Dengan memanfaatkan *unsupervised learning*, peneliti dapat memperoleh informasi baru yang sebelumnya tidak diketahui dari data yang tersedia [53].

Dalam bidang bioinformatika dan analisis data genomik, *unsupervised learning* sering digunakan untuk mengidentifikasi sub tipe kanker berdasarkan pola ekspresi gen. Sebagai contoh, pasien kanker mempunyai ciri biologis serupa dapat diklasifikasikan ke kelompok tertentu meskipun sebelumnya tidak terdapat label yang membedakan mereka. Hasil pengelompokan tersebut dapat membantu peneliti memahami heterogenitas penyakit dan menemukan biomarker baru yang berkaitan dengan perkembangan kanker [54]. Beberapa algoritma yang umum dipakai dalam *unsupervised learning* antara lain *K-Means Clustering*, *Hierarchical Clustering*, PCA, SOM [55].

2.8.3 Reinforcement Learning

Reinforcement learning merupakan metode yang mengarah pada pembelajaran melalui korelasi antara agen (*agent*) dan lingkungan (*environment*). Dalam metode ini, agen akan melakukan suatu tindakan (*action*) dan menerima *feedback* berupa penghargaan (*reward*) atau hukuman (*penalty*) dari tindakan yang dilakukan. Tujuan dari proses pembelajaran adalah menemukan strategi terbaik (*optimal policy*) yang dapat memaksimalkan total penghargaan yang diperoleh dalam jangka panjang [56].

Berbeda dari *supervised learning* yang menggunakan data yang sudah memiliki label, *reinforcement learning* tidak memerlukan data target yang telah diketahui sebelumnya. Agen akan belajar secara bertahap melalui pengalaman yang diperoleh dari berbagai tindakan yang dilakukan selama proses pelatihan. Oleh karena itu, metode ini banyak digunakan pada permasalahan yang melibatkan pengambilan keputusan secara berurutan (*sequential decision making*) [57].

Penerapan *reinforcement learning* dapat ditemukan pada berbagai bidang seperti robotika, kendaraan otonom, sistem rekomendasi, permainan komputer, serta optimasi sistem industri. Dalam bidang kesehatan, metode ini mulai digunakan untuk membantu pengambilan keputusan klinis, penentuan strategi pengobatan, dan optimasi terapi yang disesuaikan dengan kondisi pasien [58]. Beberapa algoritma yang termasuk dalam kategori *reinforcement learning* adalah *Q-Learning*, *Deep Q-Network (DQN)*, *SARSA*, dan *Policy Gradient*.

2.8.4 Hubungan Machine Learning

Machine learning memiliki kemampuan untuk mengenali pola serta hubungan antar variabel secara otomatis berdasarkan data yang tersedia. Dalam bidang kesehatan, khususnya bioinformatika, pendekatan ini banyak dimanfaatkan untuk mendukung proses diagnosis penyakit, klasifikasi kanker, identifikasi biomarker, serta prediksi prognosis pasien melalui data molekuler dan genomik [59]. Kemampuan tersebut menjadikan *machine learning* sebagai metode yang sangat sesuai untuk menganalisis data *RNA-seq* yang memiliki karakteristik berdimensi tinggi serta kompleksitas biologis yang besar.

Pada penelitian ini, *machine learning* digunakan agar dapat mengembangkan model klasifikasi yang tidak hanya mempelajari pola ekspresi gen, tetapi juga mengidentifikasi gen-gen yang berpotensi sebagai biomarker prognostik yang berhubungan dengan kondisi klinis pasien *Lung Adenocarcinoma* (LUAD), termasuk tingkat kelangsungan hidup (*survival rate*). Melalui proses pembelajaran menggunakan data berlabel, model dapat menangkap karakteristik setiap kelas dan digunakan untuk memprediksi data baru, sehingga membantu mengungkap keterkaitan antara ekspresi gen dengan outcome klinis secara lebih akurat dibandingkan metode konvensional [60].

Selain itu, penerapan *machine learning* dalam analisis ekspresi gen memungkinkan evaluasi ribuan gen secara simultan, sehingga sangat efektif dalam proses seleksi biomarker. Berbagai studi menunjukkan bahwa metode *machine learning* mampu menangani data berdimensi tinggi (*high-dimensional data*) melalui seleksi fitur dan pembelajaran pola yang relevan, sehingga dapat meningkatkan performa klasifikasi sekaligus membantu menemukan biomarker potensial yang berperan dalam kanker [61]. Oleh karena itu, pendekatan ini diharapkan dapat merancang model yang tidak sekedar melakukan klasifikasi stadium atau risiko *survival* saja, tetapi juga berkontribusi dalam identifikasi biomarker pada pasien LUAD.

Selain itu, penerapan *machine learning* dalam analisis ekspresi gen memungkinkan evaluasi ribuan gen secara simultan, sehingga sangat efektif dalam proses seleksi biomarker. Berbagai studi menunjukkan bahwa metode *machine*

learning mampu menangani data berdimensi tinggi (*high-dimensional data*) melalui seleksi fitur dan pembelajaran pola yang relevan, sehingga dapat meningkatkan performa klasifikasi sekaligus membantu menemukan biomarker potensial yang berperan dalam kanker [62]. Oleh karena itu, pendekatan ini diharapkan mampu menghasilkan model yang tidak hanya melakukan klasifikasi stadium atau risiko *survival*, tetapi juga berkontribusi dalam identifikasi biomarker pada pasien LUAD.

2.9 Support Vector Machine (SVM)

SVM adalah algoritma *supervised learning* yang dipakai untuk menyelesaikan permasalahan dalam klasifikasi dan regresi. Cara kerja SVM adalah dengan menemukan *hyperplane* yang sesuai untuk dapat memilah data ke dalam kelas yang berbeda dengan batas maksimum [63].

Pada kasus klasifikasi, tujuan utama SVM adalah menemukan batas pemisah terbaik sehingga data dari kelas yang berbeda dapat dipisahkan secara optimal. Data yang terletak paling dekat dengan *hyperplane* adalah *support vector* dan memiliki peran krusial untuk memastikan posisi batas klasifikasi [64]. Fungsi keputusan SVM dapat dituliskan seperti pada rumus 2.2 yaitu:

$$f(x) = w^T x + b \quad (2.2)$$

dengan:

- $f(x)$ = fungsi keputusan
- w = vektor bobot (*weight vector*)
- x = data masukan
- b = bias

Nilai fungsi keputusan dipakai untuk penentuan kelas pada data berdasarkan letak data terhadap *hyperplane* yang terbentuk. SVM banyak digunakan pada data ekspresi gen karena mampu menangani data berdimensi tinggi dan menghasilkan performa klasifikasi yang baik [65].

Selain digunakan pada data yang sudah dikelompokkan secara linear, SVM juga mampu menangani data yang punya pola hubungan nonlinier melalui penggunaan *kernel*. Fungsi *kernel* merupakan teknik yang dimanfaatkan untuk pemetaan data dari ruang asli ke ruang yang berdimensi tinggi sehingga data yang awalnya sulit dipisahkan dapat dipisahkan menggunakan *hyperplane* [66].

Konsep ini dikenal sebagai *kernel trick*, yaitu suatu metode yang memungkinkan proses transformasi data dilakukan tanpa harus menghitung koordinat baru secara eksplisit pada ruang berdimensi tinggi. Dengan demikian, kompleksitas komputasi dapat dikurangi sekaligus menambah kemampuan model dalam membaca pola data yang lebih rumit [67]. Secara umum, fungsi *kernel* yang digunakan pada SVM dapat dinyatakan seperti Rumus 2.3 sebagai:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (2.3)$$

dengan:

- $K(x_i, x_j)$ = fungsi *kernel*
- x_i dan x_j = pasangan data
- ϕ = fungsi transformasi ke ruang dimensi tinggi

Beberapa jenis *kernel* yang sering dipakai pada algoritma SVM antara lain sebagai berikut.

1. Linear Kernel

Linear kernel dipakai pada kondisi dimana data mampu dikelompokkan secara linear. *Kernel* ini memiliki bentuk yang sederhana dan umumnya memberikan waktu komputasi yang lebih cepat dibandingkan jenis *kernel* lainnya. Rumusnya ada pada rumus 2.4 yaitu :

$$K(x_i, x_j) = x_i^T x_j \quad (2.4)$$

2. Polynomial Kernel

Polynomial kernel digunakan untuk menangkap hubungan nonlinier antara fitur dengan memanfaatkan transformasi berbentuk polinomial. Persamaannya adalah seperti pada Rumus 2.5

$$K(x_i, x_j) = (x_i^T x_j + c)^d \quad (2.5)$$

dengan d merupakan derajat polinomial dan c merupakan konstanta.

3. Radial Basis Function (RBF)

RBF adalah salah satu *kernel* yang paling banyak dipakai karena mampu menangani berbagai pola data nonlinier yang kompleks. *Kernel* ini bekerja dengan mengukur jarak antar data menggunakan fungsi eksponensial seperti pada Rumus 2.6

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \quad (2.6)$$

dengan γ merupakan parameter yang mengatur pengaruh suatu data terhadap data lainnya.

4. Sigmoid Kernel

Sigmoid kernel memiliki karakteristik yang menyerupai fungsi aktivasi pada *Artificial Neural Network* (ANN) seperti pada Rumus 2.7.

$$K(x_i, x_j) = \tanh(\alpha x_i^T x_j + c) \quad (2.7)$$

dengan α dan c merupakan parameter yang mengontrol bentuk fungsi.

Pada penelitian yang menggunakan data RNA-seq, *kernel* RBF sering digunakan karena mampu menangani hubungan nonlinier yang kompleks pada data ekspresi gen [68]. Namun, pemilihan *kernel* yang paling sesuai tetap bergantung pada karakteristik data dan hasil evaluasi model yang diperoleh selama proses eksperimen.

2.10 LASSO

LASSO merupakan metode *featured-selection* dan regularisasi yang dimanfaatkan untuk menentukan variabel yang paling cocok dan berkaitan dengan target prediksi [68].

Pada penelitian ini, LASSO diimplementasikan menggunakan *Logistic Regression* dengan penalti L1. Pendekatan tersebut memungkinkan model tidak hanya melakukan proses klasifikasi, tetapi juga menyeleksi fitur secara otomatis berdasarkan nilai koefisien yang dihasilkan untuk setiap gen. *Logistic Regression* digunakan untuk memperkirakan probabilitas suatu sampel termasuk ke dalam kelas tertentu melalui fungsi *sigmoid*, yang dinyatakan seperti pada Rumus 2.8 yaitu:

$$P(y = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} \quad (2.8)$$

dengan:

- $P(y = 1 | x)$ = probabilitas suatu sampel termasuk ke dalam kelas positif;
- β_0 = intersep (*bias*);
- β_j = koefisien fitur ke- j ;
- x_j = nilai fitur ke- j ;
- p = jumlah fitur.

Nilai koefisien (β) diperoleh melalui proses optimasi selama pelatihan model, yaitu dengan menyesuaikan parameter agar hasil prediksi mendekati data aktual. Besarnya nilai koefisien menunjukkan tingkat kontribusi masing-masing fitur terhadap proses klasifikasi, sedangkan tanda positif atau negatif menunjukkan arah hubungan fitur terhadap kelas yang diprediksi. Semakin besar nilai absolut koefisien, semakin besar pula pengaruh fitur tersebut dalam pembentukan model.

Pada data RNA-seq, jumlah gen yang dianalisis dapat mencapai ribuan hingga puluhan ribu fitur. Tidak semua gen memiliki hubungan yang signifikan dengan stadium kanker maupun *survival rate* pasien. Oleh karena itu, diperlukan metode seleksi fitur untuk mengurangi dimensi data dan mempertahankan fitur yang paling informatif [67]. Secara matematis, fungsi optimasi LASSO dapat dituliskan seperti pada Rumus 2.9 yaitu:

$$\min \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2.9)$$

dengan:

- y_i = nilai aktual
- $\hat{y}_i$ = nilai prediksi
- β_j = koefisien variabel ke- j
- p = jumlah fitur
- λ = parameter regularisasi

Parameter λ mengontrol besarnya penalti yang diberikan pada koefisien model. Semakin besar nilai λ , semakin banyak koefisien yang diperkecil hingga bernilai nol. Fitur yang memiliki koefisien nol akan dieliminasi dari model sehingga hanya fitur yang dianggap penting yang dipertahankan [68].

Dalam penelitian kanker berbasis ekspresi gen, LASSO banyak digunakan untuk mengidentifikasi biomarker potensial dan mengurangi kompleksitas data sebelum dilakukan proses klasifikasi [68].

2.11 Evaluasi Performa Model

Evaluasi performa model adalah tahapan yang harus dilakukan pada perancangan sistem *machine learning* untuk menilai kapasitas model saat memprediksi terhadap data yang belum pernah dipakai pada proses pelatihan. Dalam penelitian berbasis data *RNA-seq*, evaluasi dilakukan sebagai bentuk validasi bahwa kemampuan generalisasi model baik terhadap data uji bukan hanya dalam mempelajari pola pada data latih [67].

Dalam penelitian klasifikasi, beberapa tolak ukur evaluasi yang umum dipakai meliputi *accuracy*, *precision*, *recall*, *F1-score*, dan *Area Under the Curve (AUC)*. Metrik *accuracy* digunakan untuk mengukur rasio prediksi yang tepat terhadap seluruh data, yang mempunyai rumus seperti pada Rumus 2.10 yaitu sebagai:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.10)$$

Sedangkan, *precision* menjelaskan tingkat ketepatan model dalam memproyeksikan kelas positif, yang dihitung dengan rumus yang ada pada Rumus 2.11 :

$$Precision = \frac{TP}{TP + FP} \quad (2.11)$$

Adapun *recall* dipakai untuk menghitung kapasitas model dalam mengidentifikasi semua sampel positif yang sebenarnya, yang dirumuskan pada Rumus 2.12 yaitu sebagai:

$$Recall = \frac{TP}{TP + FN} \quad (2.12)$$

Selanjutnya, *F1-score* dipakai untuk ukuran keseimbangan antara *precision* dan *recall*, yang dinyatakan pada Rumus 2.13 :

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.13)$$

Selain itu, AUC pada kurva ROC dipakai dalam mengukur kemampuan model untuk bisa membedakan kelas positif dan negatif di beragam nilai ambang (*threshold*). Nilai AUC yang dekat dengan 1 menunjukkan performa klasifikasi yang baik, sedangkan nilai 0.5 menandakan performa yang sama dengan tebakan acak [70].

