BAB II

TINJAUAN PUSTAKA

2.1 Kanker Payudara

Kanker payudara merupakan jenis kanker dengan prevalensi tertinggi pada wanita dan menunjukkan keragaman biologis yang kompleks. Saat ini, kanker payudara diklasifikasikan ke dalam beberapa subtipe molekuler utama, seperti *Luminal A, Luminal B, HER2-positive, dan Triple-Negative Breast Cancer (TNBC*), berdasarkan ekspresi reseptor hormon (ER, PR) serta status ekspresi HER2 dan profil genetik [8-9]. Pemahaman terhadap masing-masing subtipe sangat penting dalam menentukan pilihan terapi dan memperkirakan prognosis pasien. Kemajuan teknologi seperti *Next-Generation Sequencing* (NGS) dan analisis ekspresi gen telah memungkinkan identifikasi berbagai mutasi genetik yang berperan dalam proses karsinogenesis, termasuk mutasi pada gen *BRCA1/2*, *PIK3CA*, dan gen-gen regulator siklus sel [10]. Studi transkriptomik juga telah mengungkap adanya pola regulasi gen spesifik yang membedakan jaringan kanker dari jaringan normal, serta mengidentifikasi potensi *biomarker* baru untuk deteksi dini dan pengembangan terapi target. Pendekatan terapi modern seperti targeted therapy (misalnya *trastuzumab* untuk pasien dengan overekspresi HER2) dan imunoterapi kini dikembangkan secara lebih terarah dan berbasis molekuler [10]. Penelitian terkini berfokus pada integrasi data molekuler dan data klinis untuk mendukung pendekatan *precision medicine*, yang memungkinkan personalisasi terapi berdasarkan profil genetik dan karakteristik individu pasien.

2.2 Next Generation Sequencing (NGS)

Next Generation Sequencing (NGS) adalah teknologi pengurutan DNA/RNA generasi terbaru yang memungkinkan analisis genom secara cepat, paralel, dan berkapasitas tinggi. Berbeda dengan metode konvensional seperti Sanger sequencing, NGS dapat mengurutkan jutaan fragmen nukleotida secara simultan, sehingga efisien untuk analisis ekspresi gen, mutasi, hingga profil genomik secara menyeluruh. Dalam konteks penelitian kanker, NGS digunakan

untuk mengidentifikasi perubahan genetik dan pola ekspresi gen yang berperan dalam perkembangan dan stadium kanker. Teknologi ini menjadi dasar penting dalam bioinformatika dan pengembangan pengobatan presisi (*precision medicine*).

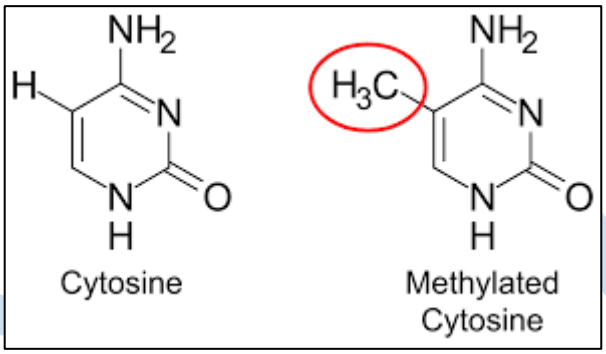
2.3 DNA Methylation

DNA Methylation adalah proses biokimia berupa penambahan gugus metil ($–CH_3$) pada basa sitosin di dalam DNA, khususnya pada posisi CpG dinukleotida [11-12] Proses ini tidak mengubah urutan DNA, tetapi memengaruhi ekspresi gen secara epigenetik, yaitu dengan mengaktifkan atau menonaktifkan gen tertentu tanpa mengubah struktur genetiknya [13]. Metilasi DNA berperan penting dalam regulasi gen, perkembangan sel, dan mekanisme pertahanan genom [14]. Pada kondisi normal, pola metilasi membantu menjaga stabilitas genom, sedangkan pada kondisi abnormal seperti kanker, terjadi hipermetilasi atau hipometilasi yang menyebabkan gangguan ekspresi gen [15]. Secara khusus, metilasi DNA terjadi pada situs-situs tertentu dalam urutan nukleotida. Yang paling umum adalah situs CpG (sitosin diikuti oleh *guanin*), yang merupakan target utama metilasi. Namun, metilasi juga dapat terjadi pada konteks non-CpG seperti CHG dan CHH, di mana H merepresentasikan A, T, atau C. Konteks ini lebih sering ditemukan pada sel punca dan jaringan tertentu, termasuk dalam kondisi patologis. Beberapa studi juga menggunakan notasi tambahan seperti CG, CH, CT, dan RS untuk mendeskripsikan lokasi atau pola spesifik metilasi dalam analisis sekuens DNA. Notasi tersebut membantu dalam mengklasifikasikan tipe metilasi berdasarkan pasangan basa atau motif yang dikenali dalam data sekuensing beresolusi tinggi.

Fitur CG merujuk pada metilasi di situs CpG (*Cytosine-phosphate-Guanine*), yang merupakan lokasi utama terjadinya perubahan epigenetik pada kanker. Metilasi abnormal di situs ini sering dikaitkan dengan inaktivasi gen penekan tumor dan deregulasi ekspresi gen, sehingga menjadikannya salah satu indikator utama dalam studi epigenetik kanker. Sementara itu, fitur CH (dengan H A, T, atau C) mencerminkan metilasi *non-CpG* yang lebih jarang ditemukan di jaringan normal, tetapi dapat meningkat dalam kondisi patologis seperti kanker, serta menggambarkan perubahan global dalam pola metilasi yang mengarah pada

deregulasi genetik. Fitur CT merepresentasikan pola metilasi yang melibatkan transisi spesifik antara basa sitosin dan timin, dan sering dihubungkan dengan mutasi atau variasi epigenetik yang memiliki relevansi terhadap prognosis pasien kanker. Adapun fitur RS (*Regional Segments*) mengacu pada pola metilasi dalam wilayah genom yang lebih luas, seperti domain regulasi atau *enhancer*, yang memungkinkan identifikasi biomarker epigenetik pada level spasial. Keempat jenis fitur ini memberikan perspektif berbeda terhadap lanskap epigenetik kanker payudara dan digunakan untuk menyusun pendekatan analisis yang lebih terfokus dan komprehensif terhadap data metilasi.

Dalam konteks kanker payudara, perubahan pola metilasi DNA di berbagai motif ini dapat digunakan sebagai *biomarker* diagnostik dan prediktif karena memengaruhi gen-gen yang terlibat dalam pertumbuhan dan penyebaran sel kanker [16]. Untuk penambahan CH3 ke DNA atau mekanisme *methylation* pada gambar 2.1



Gambar 2. 1 Mekanisme dari DNA Methylation [16]

2.4 Limma & Methylsuite

Data kompleks yang dihasilkan dari teknologi NGS memerlukan pendekatan analisis bioinformatika yang andal untuk memperoleh informasi biologis yang bermakna. Salah satu metode analisis ekspresi gen dari data NGS adalah dengan menggunakan Limma (*Linear Models for Microarray Data*), sebuah *package* yang mampu mengidentifikasi gen yang terekspresi secara berbeda antar kondisi biologis [17]. Meskipun awalnya dikembangkan untuk microarray, Limma

juga dapat diterapkan pada data *RNA-seq* dengan pendekatan transformasi log dan normalisasi tertentu.

Di sisi lain, untuk analisis data epigenetik seperti metilasi DNA, digunakan paket *MethylSuite*. *MethylSuite* merupakan kumpulan alat yang mendukung analisis data metilasi dari berbagai platform. Paket ini menyediakan fungsi-fungsi untuk praproses data, deteksi situs CpG yang mengalami perubahan tingkat metilasi, serta visualisasi hasil analisis. Dengan menggunakan alat-alat ini, peneliti dapat menggali hubungan antara regulasi genetik dan epigenetik terhadap perkembangan kanker, termasuk identifikasi *biomarker* potensial [17]

2.5 Recursive Feature Elimination (RFE)

Selain *Limma* dan *MethylSuite* yang digunakan untuk analisis ekspresi gen dan metilasi DNA, terdapat pula metode statistik lainnya yang berperan penting dalam tahap seleksi fitur, salah satunya adalah Recursive Feature Elimination (RFE) [18] . Metode ini digunakan untuk mengidentifikasi fitur-fitur paling relevan dari data berdimensi tinggi, seperti data biologis, guna meningkatkan performa model prediktif.

RFE bekerja secara iteratif dengan membangun model dan mengeliminasi fitur yang kontribusinya paling kecil terhadap prediksi, berdasarkan bobot yang diberikan oleh algoritma pembelajaran seperti SVM atau *Logistic Regression* [18]. Proses ini diulang hingga diperoleh subset fitur optimal yang memberikan performa terbaik.

Dalam penelitian biomedis, khususnya untuk klasifikasi kanker, RFE sering digunakan untuk menyaring gen atau situs metilasi penting yang berkorelasi dengan kondisi atau stadium penyakit. Dengan mengurangi fitur yang tidak relevan, RFE tidak hanya membantu meningkatkan akurasi model, tetapi juga mengurangi *overfitting* dan mempercepat waktu komputasi.

2.6 Artificial Neural Network

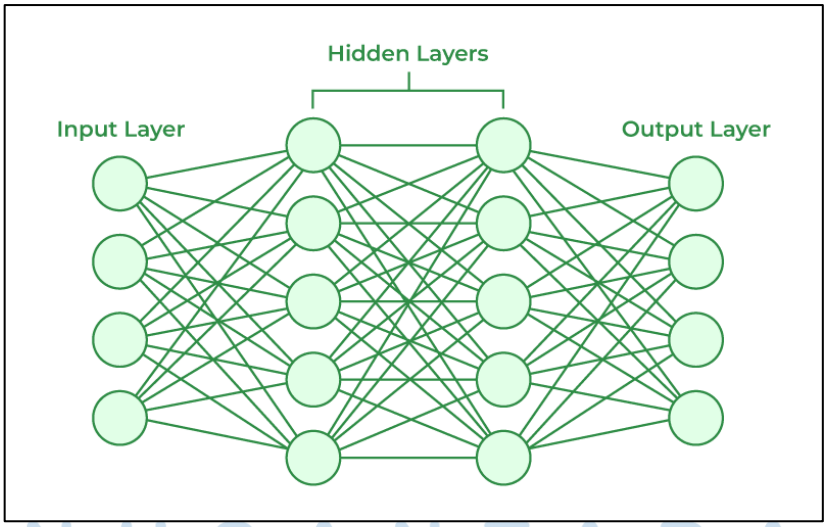
Artificial Neural Network (ANN) adalah model komputasi yang terinspirasi dari cara kerja jaringan saraf biologis di otak manusia [19-20]. ANN terdiri dari

sejumlah node atau neuron buatan yang tersusun dalam tiga lapisan utama, yaitu lapisan input, lapisan tersembunyi (hidden layer), dan lapisan output [19]. Setiap neuron dalam ANN saling terhubung dan memiliki bobot (weight) yang akan disesuaikan selama proses pelatihan untuk meminimalkan kesalahan prediksi [20].

Secara matematis, output dari sebuah neuron y dalam ANN dapat dihitung menggunakan rumus berikut:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2,1}$$

ANN mampu mempelajari hubungan kompleks antar fitur dalam data, baik yang linear maupun non-linear [21]. Dalam konteks bioinformatika dan medis, ANN sering digunakan untuk klasifikasi data seperti ekspresi gen atau metilasi DNA dalam mendeteksi atau memprediksi penyakit seperti kanker [22].



Gambar 2. 2 Struktur Artificial Neural Network (ANN) [22]

2.7 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) adalah salah satu jenis dari Artificial Neural Network (ANN) yang terdiri dari satu lapisan input, satu atau lebih lapisan

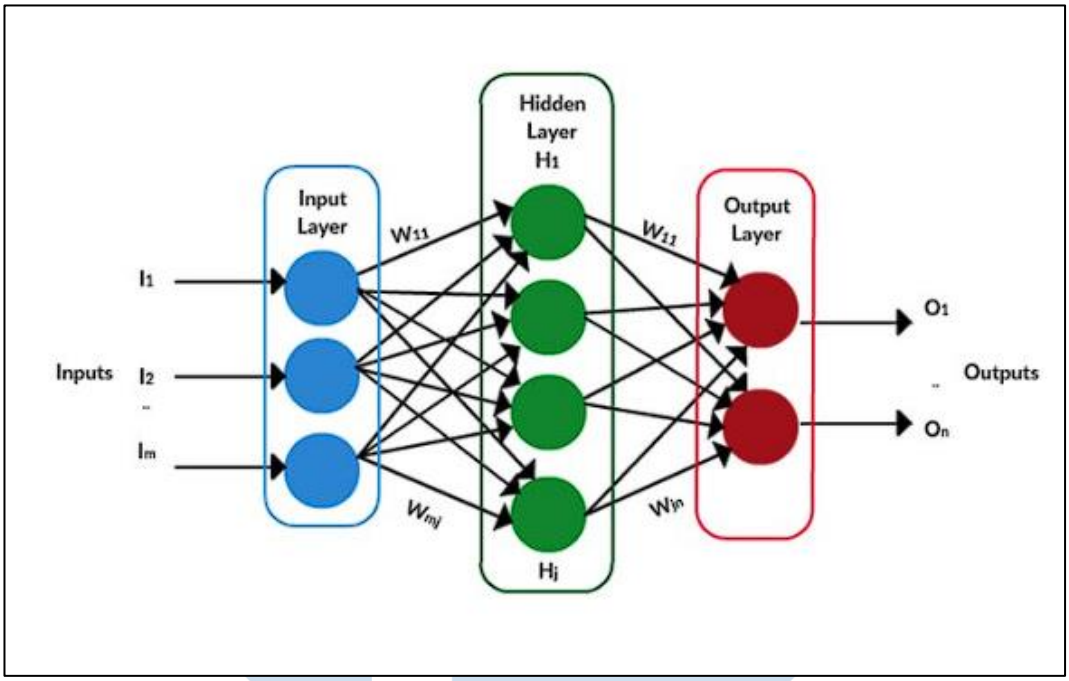
tersembunyi (*hidden layer*), dan satu lapisan output [23], [24]. Setiap neuron pada MLP terhubung secara penuh (*fully connected*) ke neuron di lapisan berikutnya [25]. MLP menggunakan fungsi aktivasi *non-linear* seperti *ReLU* atau *sigmoid* untuk memproses *input*, sehingga mampu menangani hubungan yang kompleks antar fitur [26]. Model ini dilatih menggunakan algoritma backpropagation untuk meminimalkan kesalahan prediksi melalui pembaruan bobot secara iteratif. Dalam konteks bioinformatika, MLP sering digunakan untuk membangun model klasifikasi pada data biologis seperti metilasi DNA, karena kemampuannya dalam mengidentifikasi pola yang relevan untuk prediksi penyakit seperti kanker [27].

Secara matematis, proses komputasi pada MLP dengan satu lapisan tersembunyi dapat dijelaskan sebagai berikut. Pertama, input xi diproses oleh lapisan tersembunyi menggunakan rumus:

$$h_j = f\left(\sum_{i=1}^{n} w_{ji}^{(1)} x_i + b_j^{(1)}\right) \quad (2,2)$$

Di mana $w_{ji}^{(1)}$ merupakan bobot antara input ke-i dan neuron tersembunyi ke-j dan bj adalah bias untuk neuron tersembunyi ke-j dan bj adalah bias untuk neuron tersembunyi ke-j. Output dari lapisan tersembunyi hj kemudian diteruskan ke lapisan output untuk prediksi yk sebagai berikut:

$$\widehat{y_k} = f\left(\sum_{j=1}^{m} w_{kj}^{(2)} h_j + b_k^{(2)}\right) \quad (2,3)$$



Gambar 2. 3 Arsitektur Multi-Layer Perceptron [27]

2.8 Evaluation Metrics

Dalam penelitian ini, evaluasi kinerja model klasifikasi dilakukan menggunakan beberapa metrik yang umum digunakan dalam pembelajaran mesin, yaitu *Confusion Matrix, Precision, Recall*, dan *F1-Score*. Metrik-metrik ini memberikan pemahaman yang lebih mendalam mengenai performa model, terutama dalam konteks klasifikasi data yang mungkin tidak seimbang, seperti pada kasus deteksi kanker payudara.

2.8.1 Confusion Matrix

Confusion matrix merupakan salah satu metode evaluasi yang umum digunakan dalam *machine learning*, khususnya pada model klasifikasi terawasi (*supervised classification*) [28]. Di dalam *confusion matrix*, terdapat dua jenis data utama, yaitu data aktual dan data prediksi. Data aktual adalah data yang kebenarannya sudah diketahui dan dijadikan sebagai acuan untuk menilai hasil prediksi, sementara data prediksi merupakan output dari algoritma yang digunakan

[29]. Jika disajikan dalam bentuk tabel seperti Tabel 2.1, baris dalam tabel menunjukkan kelas dari data aktual, sedangkan kolom menunjukkan kelas dari hasil prediksi. Umumnya, tabel ini berbentuk 2x2 dan mencakup empat komponen utama, yaitu:

1. True Positive (TP) – Prediksi menunjukkan hasil positif dan sesuai dengan kenyataan.
2. True Negative (TN) – Prediksi menunjukkan hasil negatif dan memang benar.
3. False Positive (FP) – Prediksi menunjukkan hasil positif namun sebenarnya salah.
4. False Negative (FN) – Prediksi menunjukkan hasil negatif namun kenyataannya salah.

Tabel 2. 1 Tabel *Confusion Matrix*

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

2.8.2 Precision

Precision adalah tingkat keakuratan model dalam memprediksi nilai positif. Nilai precision didapatkan dengan membandingkan nilai positif dengan keseluruhan prediksi positif sebuah objek. Nilai *precision* diperoleh di rumus 2,4:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2,4}$$

2.8.3 Recall

Recall menunjukkan seberapa sering model berhasil memprediksi nilai posiitf jika nilai aktualnya positif. Nilai *Recall* dapat dihitung melalui rumus 2,5:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2,5}$$

2.8.4 F1-Score

F1-Score merupakan salah satu metrik evaluasi yang menghitung keseimbangan antara *precision* dan *recall*, nilai *F1-Score* dapat didapatkan melalui rumus 2,6:

$$\text{F1 Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2,6}$$