BAB II

TINJAUAN PUSTAKA

2.1. Kanker Prostat

Kanker prostat merupakan penyakit kanker organ yang paling umum ditemui pada pria, terutama pada usia lanjut dan menunjukkan variasi biologi signifikan [23]. Heterogenitas ini terlihat dari keberadaan berbagai subtype dan pola ekspresi gen yang berbeda pada kanker prostat yang mempengaruhi jalannya penyakit dan respon terhadap pengobatan. Kanker prostat pada umunya diklasifikasikan berdasarkan skor *gleason* dan status androgen receptor (AR). Skor *gleason* digunakan untuk menilai tingkat keganasan tunor dimana skor yang rendah (<= 6) menandakan kanker yang berkembang secara lambat, sedangkan skor tinggi (8-10) menandakan kanker yang bertumbuh lebih agresif [24]. Sementara itu, ekspresi AR yang tinggi mengindikasikan tingkat agresivitas kanker dan menjadi indicator penting dalam mennetukan respon terhadap terapi hormon.

Pada tingkat molekuler, mutase genetic seperti TMPRSS2-ERG, PTEN, dan SPOP telah diindentifikasi sebagai faktor penting dalam karsinogensis kanker prostat [25], [26], [27]. Mutasi-mutasi ini tidak hanya memengaruhi jalannya penyakit, namun juga membuka peluang untuk memahami pola molekuler spesifik yang membedakan kanker prostat dari jaringan normal. Contoh, fusion gen TMPRSS2-ERG sering dikaitkan dengan aktivasi jalur transkripsi abnormal, sementara hilangnya fungsi gen PTEN dapat menyebabkan disregulasi pertumbuhan sel, dan mutasi gen SPOP memengaruhi degradasi protein regulator. Analisis terhadap data mutasi ini memungkinkan pembuatan profil subtipe molekuler, pola ekspresi gen, serta jalur sinyal yang dominan pada masing-masing kelompok pasien. Proses penemuan pola ekspresi gen spesifik yang membedakan jaringan kanker dengan normal telah berkembang pesat akibat kemajuan dalam teknologi *sequencing* dan analisis transkriptomik serta mengungkap berbagai biomarker baru untuk diagnosis kanker dan terapi yang lebih spesifik [28].

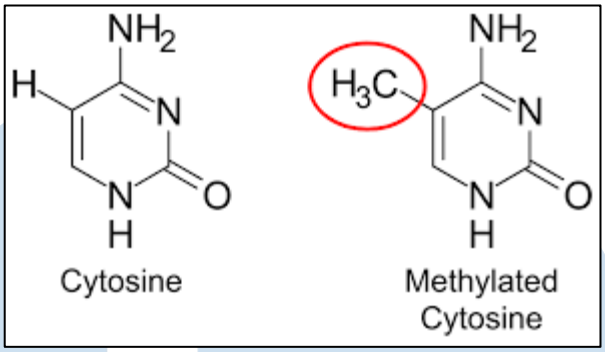
2.2. Next-Generation Sequencing

Next-Generation Sequencing atau yang biasa disingkat NGS merupakan sebuah metode yang digunakan untuk membaca *sequence* DNA/RNA secara cepat dan terjangkau [29]. Dibandingkan oleh metode *sequencing* tradisional seperti *Sanger sequeincing* yang membutuhkan waktu beberapa tahun dan biaya yang mencapai miliaran, NGS dapat menganalisis jutaan fragmen DNA atau RNA secara bersamaan dan pararel. Kemajuan metode *sequencing* ini memungkinkan NGS untuk mempersingkat waktu dan biaya yang dibutuhkan untuk *sequencing* sebuah genom manusia menjadi dalam hitungan hari dan dengan harga dibawah $1000 [30]. Muncul pada sekitar tahun 2000, NGS berperan besar dalam mendorong kemampuan manusia untuk mengungkap lebih dalam mengenai penyakit pada tingkat individu. NGS digunakan oleh peneliti dan dokter pada rumah sakit atau laboratorium untuk mempelajari penyakit, sifat-sifat dari gen dan mengembangkan pengobatan yang bersifat personal/individual.

2.3. Metilasi DNA

Metilasi DNA adalah proses penambahan gugus metil pada molekul DNA, biasanya pada posisi CpG dinukleotida yang merupakan pasangan antara sitosin dan guanin dalam urutan DNA (Gambar 1) [31]. Proses ini merupakan salah satu bentuk modifikasi epigenetik yang dapat diturunkan, bersifat reversible, dan berperan penting dalam pengaturan ekspresi gen.

Dalam konteks kanker, pola metilasi DNA sering mengalami perubahan yang pada umumnya dibagi jadi dua jenis; *hypermethylation* dan *hypomethylation* [32]. *Hypermethylation* biasanya dikaitkan dengan perkembangan tumor karena terjadi pada area promotor gen penekan tumor yang dapat membuat gen tersebut menjadi tidak aktif. Sebaliknya, *hypomethylation* biasanya terjadi secara luas di seluruh genom dan jumlahnya meningkat seiring dengan perkembangan atau tingkat keganasan tumor, meskipun hal ini bervariasi tergantung jenis tumor dan karakteristik individu [33]. Selain itu, *hypomethylation* juga berkait kepada stadium kanker, ukuran tumor, dan derajat histologis pada berbagai jenis kanker.



Gambar 1. Visualisasi Metilasi DNA dari Cytosine.

2.4. Differentially Methylated Positions

Differentially Methylated Positions (DMPs) adalah dinukleotida CpG tunggal yang menunjukkan perbedaan tingkat metilasi yang signifikan secara statistik antara dua atau lebih kelompok biologis misalnya, sampel dari pasien dengan kanker dan kontrol sehat, atau antara stadium kanker yang berbeda [34]. Analisis identifikasi DMPs merupakan salah satu langkah utama dalam analisis epigenetik karena dapat mengungkapkan perubahan regulasi genetik yang relevan dengan kondisi biologis atau penyakit tertentu. Setiap DMP diukur menggunakan beta value, yaitu rasio intensitas sinyal metilasi terhadap total sinyal (metilasi + non-metilasi), dengan rentang nilai dari 0 (tidak termetilasi) hingga 1 (sepenuhnya termetilasi) [35]. Nilai tingkat metilasi ini lalu dibandingkan antar kelompok, dan uji statistic seperti t-test, ANOVA, atau model linear digunakan untuk menentukan apakah perbedaan tingkat metilasi pada posisi tersebut signifikan.

2.5. Recursive Feature Elimitation

Recursive Feature Elimination (RFE) adalah metode seleksi fitur yang melakukan eliminasi berulang-ulang terhadap fitur-fitur paling tidak informatif berdasarkan bobot atau skala kepentingan yang diberikan oleh suatu estimator [36]. Metode seleksi fitur ini efektif untuk mereduksi dimensi data seiring dengan mempertahankan fitur-fitur paling relevan.

Pada awal proses, estimator dilatih menggunakan seluruh fitur dalam dataset. Setelah model melalui proses training, setiap fitur akan diberikan skor berdasarkan bobot (untuk model linear) atau nilai kepentingan fitur (untuk model berbasis tree).

Skor ini digunakan untuk mengidentifikasi fitur mana yang paling tidak berkontribusi terhadap hasil prediksi model. Fitur dengan skor terendah kemudian dieliminasi dari dataset. Proses pelatihan ulang dilakukan pada dataset dengan fitur yang sudah dikurangi, dan langkah eliminasi diulang secara bertahap sampai jumlah fitur tersisa mencapai target tertentu atau hingga tidak ada fitur lagi yang bisa dihapus. Secara umum, fungsi pemilihan fitur RFE dapat direpresentasikan sebagai pada pseudocode berikut [37]:

Algoritma 2.1: Recursive Feature Elimination

Input:

Training dataset

Set of Features

Ranking method

Output:

Final ordered list of selected features

Steps:

1. Repeat for each step from 1 to the total number of features:
2. Use the ranking method to evaluate and rank the current set of features
3. Identify the feature with least importance
4. Add that feature to the final ranking list
5. Remove that feature from the current st of features

Efektivitas RFE bergantung pada jenis dataset. Pada dataset dengan fitur yang sangat berkorelasi, RFE dapat menghapus salah satu dari fitur yang sebenarnya saling melengkapi [38]. Selain itu, dataset dengan distribusi label tidak seimbang dapat mempengaruhi skor kepentingan yang dihasilkan oleh estimator, sehingga penting untuk menggunakan metode stratifikasi sampel atau penyeimbangan kelas sebelum menjalankan RFE.

2.6. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) merupakan jenis jaringan saraf tiruan dengan tipe *feedforward* yang banyak digunakan untuk tugas klasifikasi dan regresi [39]. MLP terdiri dari sebuah *input layer*, satu atau lebih *hidden layer*, dan *output layer*.

MLP belajar dengan menyesuaikan bobot yang dikenakan pada neuron untuk meminimalisir perbedaan antara hasil prediksi dengan nilai actual, proses ini biasanya menggunakan teknik *Backpropagation*.

Algoritma 2.2 *backpropagation* yang menghitung turunan error terhadap setiap bobot dan memperbarui bobot memiliki rumus:

$$\omega_{ij}^{(t+1)} = \omega_{ij}^{(t)} - \eta \frac{\partial E}{\partial \omega ij} \quad (2.1)$$

di mana ∂ adalah learning rate dan E adalah fungsi error.

Algoritma 2.2: Back Propagation

Input:

Training dataset
Initial weights for all neurons
Learning Rate
Number of epochs
Activation functions
Loss function

Output:

Updated weights after training

Steps:

1. Repeat for each epoch:
2. Perform forward propagation to compute prediction for each layer in the network:
    - Calculate weighted sum of inputs
    - Apply activation function to get output of the neuron
3. Compute error between prediction and true label using loss function
4. Perform backward propagation to update weights
    - Calculate the error gradient for each output neuron
    - Propagate the error back to previous layers
    - Compute gradients for each weight based on the error
    - Update each weight by subtracting the product of the learning rate and its gradient
5. End training after all epochs are complete

MLP terdiri dari beberapa lapisan *neuron* (Gambar 2) diantaranya adalah; lapisan penerima (*input layer*) memiliki jumlah neuron yang bervariasi akibat

ketergantungannya kepada data yang akan dimasukan ke MLP sebagai fitur. Lapisan penerima berfungsi untuk menerima dan meneruskan data mentah yang digunakan sebagai fitur ke lapisan selanjutnya [40]. Lapisan tersembunyi (*hidden layer*) berperan untuk melakukan pembelajaran atau komputasi guna untuk mepelajari pola-pola yang terdapat pada data dengan menggunakan fungsi aktivasi (*activation function*).

Untuk setiap neuron di hidden layer pertama dihitung dengan rumus berikut:

$$h_{ij} = f\left(\sum_{k=1}^{n_0} \omega_{kj} x_k\right) \tag{2.2}$$

dimana x_k merupakan nilai fitur yang di-*input*, ω_{kj} adalah bobot input neuron k kepada *hidden* neuron j, dan n_0 adalah jumlah dari *input* neuron.

Output setiap neuron pada *hidden layer* selanjutnya dihitung dengan rumus berikut:

$$h_{ij} = f\left(\sum_{k=1}^{n_i-1} \omega_{kj}^{(i)} h_{(i-1)}k\right) \tag{2.3}$$

dimana f adalah fungsi aktivasi, $\omega_{kj}^{(i)}$ adalah bobot tantara neuron ke-k di lapisan sebelumnya ke neuron ke-j di lapisan i, dan $h_{(i-1)}k$ adalah *output* dari neuron ke-k di lapisan sebelumnya.

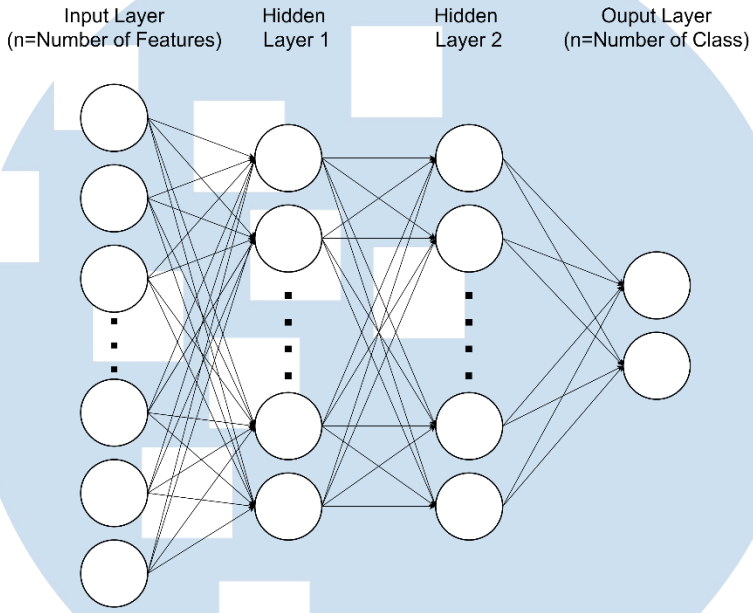
Hasil akhir (*output layer*) jaringan akan dihitung dengan:

$$y_l = f\left(\sum_{k=1}^{nN} \omega_{kl}^{(N+1)} h_{Nk}\right) \tag{2.4}$$

dimana h_{Nk} adalah output dari *hidden layer* terakhir dan $\omega_{kl}^{(N+1)}$ adalah bobot dari neuron ke-k di *hidden layer* terakhir ke output neuron ke-i.

Lapisan keluaran (*output layer*) bertugas untuk menghasilkan hasil akhir atau prediksi dari MLP, neuron pada layer ini bergantung pada kebutuhan klasifikasi jika regresi dan klasfiikasi biner hanya memerlukan satu neuron, sementara klasfikasi multikelas memerlukan satu neuron untuk setiap kelas. Semua lapisan

dari MLP bekerja sama mempelajari pola kompleks dalam data melalui banyak iterasi yang kemudian menghasilkan luaran berupa prediksi kelas.



Gambar 2. Kerangka Multi-Layer Perceptron.

2.7. Metrik Evaluasi

Secara umum, mengevaluasi model ML ataupun DL melibatkan penilaian terhadap seberapa baik model tersebut dapat memberdakan antara kelas-kelas yang ada dengan menggunakan beberapa metrik. Metrik ini membantu mengidentifikasi kelebihan dan kelemahan model, terutama pada performa model dalam menangani ketidakseimbangan kelas atau kasus diluar batas. Salah satu metode yang umum digunakan adalah dengan menggunakan *classification report*, yang menyediakan metrik rinci seperti *accuracy, precision, recall,* dan *F1-score*. Metrik-metrik ini menawarkan visualisasi terhadap performa terutama dalam masalah klasifikasi biner dimana skor akurasi yang sederhana dapat tidak mencerminkan performa actual model.

- Precision = $\frac{TP}{(TP + FP)}$
- Recall = $\frac{TP}{(TP + FN)}$
- F1-score = $\frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$

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

Selain dari *classification report*, AUC (Area Under the Curve) dari kurva ROC (Receiver Operating Characteristic) sering digunakan sebagai metrik pengukur kemampuan model dalam membedakan antar kelas positif dan negative di semua nilai batas atas. Nilai AUC berkisar antara 0 sampai 1 dimana nilai 0,5 menunjukan bahwa model menghasilkan prediksi secara acak sementara nilai 1 menunjukan bahwa model menghasilkan prediksi yang sempurna terhadap semua label. AUC penting ketika bekerja dengan dataset yang tidak seimbang (*imbalanced*), karena AUC berfokus kepada kinerja klasifikasi model daripada hanya mengandalkan distribusi kelas.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA