BAB 2
LANDASAN TEORI

2.1 Kanker Paru

Kanker paru didefinisikan sebagai kondisi yang ditandai dengan pertumbuhan sel yang tidak terkendali (*uncontrolled cell growth*) pada jaringan paru, umumnya bermula dari sel-sel epitel yang melapisi saluran pernapasan atau bronkus [40]. Secara molekuler, pertumbuhan abnormal ini dipicu oleh kerusakan *deoxyribonucleic acid* (DNA), yang mencakup inaktivasi gen penekan tumor (*tumor suppressor genes*) serta aktivasi proto-onkogen menjadi onkogen [40]. Seiring perkembangannya, sel-sel ini membentuk massa tumor yang tidak hanya mengganggu fungsi vital sistem pernapasan, tetapi juga memiliki kemampuan untuk menginvasi jaringan di sekitarnya dan menyebar ke organ tubuh lainnya (*metastasis*), yang pada akhirnya memperburuk kondisi klinis penderita [40], [41].

Data tahun 2022 menunjukkan bahwa kanker paru menyumbang sekitar 2,48 juta kasus baru serta menyebabkan sekitar 1,81 juta kematian [1]. Tanpa adanya intervensi terapeutik baru yang lebih efektif, angka morbiditas dan mortalitas ini diproyeksikan akan terus melonjak secara drastis hingga tahun 2050 [2]. Tingginya angka kematian ini berkaitan erat dengan kesulitan deteksi dini, mengingat kanker paru sering berkembang tanpa gejala pada tahap awal sehingga sebagian besar pasien baru terdiagnosis pada stadium lanjut, dengan proporsi mencapai sekitar 84% [42]. Keterlambatan diagnosis tersebut berdampak langsung pada rendahnya kelangsungan hidup pasien, yang tercermin dari tingkat kelangsungan hidup lima tahun (*five-year survival rate*) yang masih relatif rendah, yaitu berkisar antara 17,4% hingga 27% [42], [43]. Meskipun perkembangan terapi target dan imunoterapi telah meningkatkan pilihan pengobatan bagi pasien stadium lanjut, resistensi terhadap pengobatan dan metastasis tetap menjadi tantangan klinis utama dalam upaya menurunkan angka kematian kanker paru [6], [43]. Oleh karena itu, diperlukan eksplorasi kandidat obat antikanker baru yang mampu mengintervensi target molekuler secara lebih efektif guna mengatasi mekanisme resistensi tersebut.

2.2 Tanaman Herbal

Tanaman herbal merupakan tanaman yang mengandung senyawa fitokimia dengan aktivitas biologis yang dapat dimanfaatkan untuk tujuan pengobatan maupun pencegahan penyakit [44]. Kondisi geologis dan iklim tropis Indonesia menghasilkan variasi genetik tanaman yang melimpah, menjadikan endemisitas flora Nusantara sebagai sumber utama yang sangat menjanjikan dalam pencarian agen terapeutik modern [8], [9]. Kekayaan botani ini secara historis telah diintegrasikan dalam pengobatan tradisional masyarakat Nusantara

yang dikenal sebagai *Jamu* [12]. Saat ini, tradisi tersebut perlahan disaintifikasi ke dalam sistem kesehatan formal melalui pendekatan farmakologi modern guna menemukan agen terapeutik yang terstandardisasi [10], [27].

Dalam bidang onkologi, berbagai studi *in vitro* mengonfirmasi bahwa ekstrak tanaman lokal memiliki kemampuan sitotoksik dan antiproliferatif yang selektif terhadap sel kanker manusia [45]. Dibandingkan dengan kemoterapi konvensional yang sering memicu toksisitas sistemik dan resistensi obat, agen bersumber alam menawarkan profil keamanan yang jauh lebih baik terhadap sel sehat [45], [46]. Kemampuan intervensi target molekuler yang presisi dan tingkat toksisitas yang rendah ini menjadikan tanaman herbal sangat potensial untuk dikembangkan sebagai agen kemopreventif sekaligus terapi pendamping (*co-chemotherapy*) guna memaksimalkan efikasi dan meminimalkan efek samping dari kemoterapi standar [46].

2.3 Senyawa Bioaktif

Senyawa bioaktif umumnya didefinisikan sebagai molekul non-nutrisi yang diekstrak dari organisme hidup, khususnya tanaman, dengan aktivitas farmakologis yang signifikan [44]. Dalam konteks botani, senyawa ini secara kolektif dikenal sebagai metabolit sekunder dan mencakup kelompok alkaloid, flavonoid, terpenoid, serta senyawa fenolik [47], [48]. Berbeda dengan metabolit primer yang berperan dalam proses pertumbuhan dan metabolisme dasar, metabolit sekunder disintesis oleh tanaman sebagai mekanisme pertahanan terhadap tekanan lingkungan, baik abiotik maupun biotik [47], [48], [49]. Metabolit ini diproduksi untuk mendukung adaptasi ekologis tanaman, termasuk perlindungan terhadap patogen dan herbivora, serta respons terhadap perubahan kondisi lingkungan yang ekstrem [50]. Secara farmakologis, berbagai golongan metabolit sekunder tersebut terbukti memberikan efek terapeutik yang penting bagi manusia [51]. Senyawa fitokimia ini telah dipelajari secara ekstensif karena kemampuannya memodulasi jalur biologis dan menawarkan potensi aplikasi medis yang luas, mulai dari aktivitas antioksidan dan antiinflamasi hingga berperan sebagai agen antikanker [44], [51].

Pada pengembangan obat modern, berbagai kelompok senyawa turunan alam ini menawarkan keunggulan unik berupa keragaman struktur kimia (*chemical space*) dan kompleksitas stereokimia yang jauh lebih luas dibandingkan senyawa sintetik, sehingga meningkatkan peluang ditemukannya kandidat obat baru yang mampu berinteraksi dengan target protein penyakit yang kompleks [31], [52]. Pemanfaatan potensi ini kini didukung oleh ketersediaan berbagai basis data publik yang mengkurasi ribuan struktur senyawa bioaktif tanaman. Data terstruktur ini menjadi elemen kunci dalam pendekatan *Computer-Aided Drug Design* (CADD), yang memungkinkan melakukan penapisan virtual skala besar secara efisien [53]. Penapisan virtual ini menjadi semakin mutakhir ketika diintegrasikan

dengan kerangka *machine learning* yang didesain khusus untuk mengekstraksi dan memetakan interaksi protein di dalam ruang kimiawi produk alami tersebut [31].

2.4 Machine Learning

Machine learning (ML) adalah proses otomatis yang memungkinkan sistem komputasi untuk mengidentifikasi dan mempelajari pola dalam data untuk melakukan tugas-tugas spesifik dan kompleks tanpa diprogram secara eksplisit [54], [55]. Sebagai bagian fundamental dari *artificial intelligence*, algoritma ML umumnya diklasifikasikan ke dalam tiga paradigma utama, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [21], [23]. Sementara *unsupervised learning* berfokus pada pengelompokan data tanpa label dan *reinforcement learning* pada pengambilan keputusan berbasis *reward*, *supervised learning* telah terbukti menjadi kerangka kerja utama yang mendominasi pemodelan farmakologi prediktif [21], [55]. Dalam pendekatan *supervised learning* ini, algoritma dilatih menggunakan himpunan data yang telah teranotasi, seperti pasangan interaksi senyawa-target yang diketahui, untuk membangun fungsi pemetaan prediktif. Fungsi ini kemudian secara efisien digunakan untuk mengklasifikasikan dan memprediksi aktivitas farmakologis dari kandidat molekul baru yang belum diobservasi sebelumnya [54], [55].

Model statistik parametrik tradisional sering kali kesulitan memproses data yang sangat tidak terstruktur serta interaksi non-linear yang kompleks pada sistem biologis [21]. Untuk mengatasi keterbatasan tersebut, algoritma *ensemble* berbasis pohon (*tree-based*) dan kerangka kerja *machine learning* tingkat lanjut telah muncul sebagai alternatif yang ampuh dalam bidang keminformatika dan bioinformatika [56], [57]. Algoritma tersebut beroperasi dengan membagi data secara berulang-ulang berdasarkan karakteristik prediktif ke dalam kelompok yang semakin homogen, sehingga menjadikannya sangat kuat untuk menangani ruang fitur berdimensi tinggi. Berbagai studi secara konsisten menunjukkan keunggulan algoritma seperti *Random Forest* dan kerangka kerja *gradient boosting* tingkat lanjut (misalnya, XGBoost) dalam mengidentifikasi biomarker penyakit kompleks dan memprediksi hasil klinis di berbagai bidang medis [58], [59], [60].

2.4.1 Random Forest

Random Forest (RF) adalah algoritma *supervised learning* berbasis *ensemble* yang membangun sekumpulan pohon keputusan yang tidak dipangkas (*unpruned decision trees*) selama fase pelatihan untuk menghasilkan prediksi konsensus [54], [61]. RF diperkenalkan oleh Breiman untuk mengatasi keterbatasan *overfitting* dari pohon keputusan standar dengan memanfaatkan *bootstrap aggregating* (*bagging*) bersamaan dengan *random feature*

selection [61]. Dengan menggunakan keacakan (*randomness*) pada sampel pelatihan dan subset variabel yang dievaluasi di setiap *node*, RF secara signifikan meningkatkan generalisasi dan ketahanan (*robustness*) model terhadap *noise* atau ketidakseimbangan kelas [62].

Algorithm 2.1 Basic Random Forest

Require: Training set D, number of trees T, feature subset size m
Ensure: Ensemble model RF

$RF \leftarrow \emptyset$
for $t = 1$ to T **do**
    $D_t \leftarrow$ Generate bootstrap sample from D (sampling with replacement)
    Initialize tree $Tree_t$ with root node containing D_t
    while stopping criteria is not met **do**
        Select m features randomly from all available features
        Find the best feature and split point among the m features (e.g., minimizing Gini Impurity)
        Split the node into child nodes
    end while
    Add $Tree_t$ to the ensemble: $RF \leftarrow RF \cup \{Tree_t\}$
end for
return RF

Selama proses pembentukan setiap pohon, algoritma harus menentukan pemisahan (*split*) yang optimal pada setiap *node*. Untuk tugas klasifikasi biner, seperti memprediksi apakah suatu senyawa dan target akan berinteraksi, hal ini umumnya dicapai dengan meminimalkan *Gini impurity*. *Gini impurity*, $i(\tau)$, mengevaluasi seberapa baik suatu pemisahan membagi sampel pada *node* τ tertentu, dan secara matematis didefinisikan sebagai:

$$i(\tau) = 1 - p_p^2 - p_n^2 \tag{2.1}$$

di mana p_p mewakili proporsi sampel positif (pasangan yang berinteraksi) and p_n mewakili proporsi sampel negatif (pasangan yang tidak berinteraksi) yang mencapai *node* tersebut. Algoritma melakukan pencarian secara menyeluruh (*exhaustive search*) pada subset fitur yang dipilih secara acak untuk menemukan ambang batas fitur (*feature threshold*) yang memaksimalkan penurunan *impurity* ini [63].

Setelah seluruh pohon terbentuk, klasifikasi akhir ditentukan dengan menggabungkan prediksi dari semua pohon individu melalui pemungutan suara terbanyak (*majority vote*) [61]. Lebih lanjut, algoritma ini memiliki kemampuan interpretasi bawaan untuk menilai tingkat kepentingan relatif (*relative importance*) dari setiap fitur dengan

mengukur total penurunan *node impurity* di seluruh pohon, sehingga memberikan wawasan struktural yang transparan [29], [63].

2.4.2 Advanced Gradient Boosting: XGBoost and LightGBM

Algoritma *gradient boosting* menggunakan pendekatan *ensemble* sekuensial di mana setiap pohon keputusan baru dioptimalkan secara khusus untuk meminimalkan kesalahan residual dari *ensemble* sebelumnya [28]. Di antara implementasi teknik ini yang paling canggih dan efektif dalam farmakologi komputasional adalah *eXtreme Gradient Boosting* (XGBoost) dan *Light Gradient Boosting Machine* (LightGBM) [31].

Algorithm 2.2 Basic eXtreme Gradient Boosting (XGBoost)

Require: Training data D, number of iterations T, learning rate η
Ensure: Ensemble model $F(x)$
1: Initialize $F_0(x)$ with a constant value
2: **for** $t = 1$ to T **do**
3: Compute first-order gradients (g_i) and second-order hessians (h_i) for all instances
4: Initialize a new decision tree f_t
5: **for** each node in f_t **do**
6: Evaluate potential splits across features
7: Choose the split that maximizes the Regularized Objective Gain
8: **end for**
9: Update the ensemble model: $F_t(x) \leftarrow F_{t-1}(x) + \eta f_t(x)$
10: **end for**
11: **return** $F_T(x)$

XGBoost merupakan sistem *boosting* berskalabilitas tinggi (*highly scalable*) yang menyempurnakan *standard gradient boosting* dengan memanfaatkan aproksimasi orde kedua (*second-order approximation*) untuk mengoptimalkan *loss function* secara cepat, sekaligus menerapkan regularisasi ketat (*strict regularization*) untuk memenalti kompleksitas model [64]. Pendorong matematis utama dari XGBoost adalah fungsi objektif yang diregularisasi (*regularized objective function*), yang menentukan bagaimana model belajar pada setiap langkah sekuensial. Pada iterasi ke-t, fungsi objektif yang disederhanakan (*simplified objective function*) $\tilde{\mathscr{L}}^{(t)}$ didefinisikan sebagai:

$$\tilde{\mathscr{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.2}$$

di mana g_i dan h_i mewakili statistik gradien orde pertama dan orde kedua pada *loss function*, $f_t(x_i)$ adalah prediksi dari pohon baru untuk instans ke-i, dan $\Omega(f_t)$ adalah suku

regularisasi (*regularization term*) yang mengontrol kompleksitas pohon untuk mencegah *overfitting* [64]. Fondasi matematis yang kuat (*robust*) ini memungkinkan XGBoost secara efektif menangkap dependensi non-linear yang kompleks dalam dataset berdimensi tinggi.

LightGBM dibangun berdasarkan prinsip-prinsip *gradient boosting* namun dioptimalkan secara arsitektur untuk pelatihan yang terdistribusi dan sangat efisien pada dataset berskala masif. Seperti yang diperkenalkan oleh Ke et al., implementasi *boosting* tradisional mengharuskan pemindaian seluruh instans data untuk mengestimasi *information gain* pada setiap fitur, yang mana sangat mahal secara komputasi [65]. Untuk mengatasi hal ini, kerangka kerja memperkenalkan dua teknik baru, yaitu *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB). GOSS mempertahankan instans data dengan gradien besar (yang berkontribusi lebih banyak terhadap *information gain*) dan melakukan pengambilan sampel secara acak pada instans dengan gradien kecil, sehingga memungkinkan algoritma untuk mencapai estimasi *gain* yang sangat akurat dengan ukuran data yang berkurang secara signifikan. Dikombinasikan dengan strategi pertumbuhan pohon *leaf-wise*, arsitektur ini menjadikan LightGBM sangat cocok untuk menangani ruang fitur berdimensi tinggi yang renggang (*sparse*), seperti yang umum ditemukan pada *2D molecular fingerprinting* dalam tugas penemuan obat (*drug discovery*) [56], [65].

Algorithm 2.3 Basic LightGBM (GOSS and Leaf-wise Growth)

Require: Training data D, iterations T, top gradient ratio a, small gradient ratio b
Ensure: Ensemble model $F(x)$

for $t = 1$ to T **do**
  Compute absolute gradients for all instances in D
  Sort instances by gradient magnitude in descending order
  $TopSet \leftarrow$ Select top $a\%$ of instances (large gradients)
  $RandSet \leftarrow$ Randomly sample $b\%$ of the remaining instances (small gradients)
  Apply weight multiplier $\frac{1-a}{b}$ to gradients in $RandSet$ to maintain data distribution
  $Subset \leftarrow TopSet \cup RandSet$
  Build tree f_t using $Subset$ with a Leaf-wise growth strategy (split node with max loss)
  Update the ensemble model: $F_t(x) \leftarrow F_{t-1}(x) + f_t(x)$
end for
return $F_T(x)$

2.4.3 Cascade Deep Forest

Cascade Deep Forest (gcForest), yang diperkenalkan oleh Zhou dan Feng, adalah kerangka kerja pembelajaran *ensemble* tingkat lanjut yang memanfaatkan arsitektur berlapis

(*layered architecture*) yang dalam yang seluruhnya terdiri dari hutan pohon keputusan [66]. Arsitektur ini mempertahankan kemampuan *hierarchical representation learning* dari model *deep learning* namun tetap mempertahankan kemudahan pelatihan, penentuan kompleksitas secara otomatis, dan interpretabilitas teoretis yang melekat pada algoritma berbasis pohon, sehingga sangat cocok untuk memproses kumpulan data keminformatika tabular dan berdimensi tinggi yang umum dalam alur kerja penemuan obat [24], [66].

Mekanisme utama dari *Cascade Deep Forest* adalah struktur pemrosesan lapis demi lapis (*layer-by-layer processing structure*). Setiap tingkat dalam *cascade* terdiri dari *ensemble* beberapa *forest* (biasanya merupakan campuran dari *completely-random tree forests* dan *standard random forests*) untuk mendorong keberagaman (*diversity*). Setiap *forest* memproses input dan menghasilkan estimasi distribusi kelas, yang kemudian membentuk sebuah *class vector*. Untuk memungkinkan *hierarchical representation learning*, input ke lapisan berikutnya bukan hanya output dari lapisan sebelumnya, melainkan sebuah *augmented vector*. Secara matematis, input untuk tingkat *cascade* ke-$(t+1)$, yang dinotasikan sebagai $\mathbf{H}^{(t+1)}$, didefinisikan sebagai:

$$\mathbf{H}^{(t+1)} = \left[\mathbf{x}, \mathbf{h}_1^{(t)}, \mathbf{h}_2^{(t)}, \ldots, \mathbf{h}_K^{(t)}\right] \tag{2.3}$$

di mana $\mathbf{x}$ adalah *raw input feature vector* asli, dan $\mathbf{h}_k^{(t)}$ mewakili *class probability vector* yang dihasilkan oleh *forest* ke-k pada tingkat sebelumnya t. Untuk mengurangi risiko *overfitting*, vektor kelas ini dihasilkan menggunakan *k-fold cross-validation* [66].

Dengan terus menggabungkan (*concatenating*) fitur asli dengan distribusi probabilitas yang dipelajari, model secara efektif mengekstraksi pola hierarkis non-linear yang kompleks tanpa mengalami *vanishing gradients* yang umum terjadi pada DNNs. Selanjutnya, algoritma secara adaptif menentukan kompleksitas modelnya sendiri dengan menghentikan pembuatan tingkat *cascade* baru ketika performa validasi tidak lagi meningkat. Dalam konteks memprediksi interaksi obat-target untuk pengobatan kanker paru, arsitektur ini sangat mudah beradaptasi. Algoritma ini secara efisien memproses deskriptor tabular berdimensi tinggi yang diekstraksi dari senyawa alami, sehingga menghasilkan prediksi ikatan (*binding predictions*) yang sangat akurat meskipun skala data pelatihan terbatas [32], [66].

Algorithm 2.4 Basic Cascade Deep Forest

Require: Raw input features $\mathbf{x}$, maximum layers M

Ensure: Cascade Forest Model

Layer index $l \leftarrow 1$
Cascade Input $\mathbf{H}^{(1)} \leftarrow \mathbf{x}$
$BestScore \leftarrow 0$
while $l \leq M$ **do**
    Train a variety of random forests (e.g., standard and completely-random) using $\mathbf{H}^{(l)}$
    Generate class probability vectors $\mathbf{h}_k^{(l)}$ from the forests via cross-validation
    Concatenate raw features with new probabilities: $\mathbf{H}^{(l+1)} \leftarrow [\mathbf{x}, \mathbf{h}_k^{(l)}]$
    Evaluate validation performance of layer $l \rightarrow Score^{(l)}$
    if $Score^{(l)} > BestScore$ **then**
        $BestScore \leftarrow Score^{(l)}$
        Save layer l structure
        $l \leftarrow l + 1$
    else
        break (Early stopping triggered, performance did not improve)
    end if
end while
return Trained model up to optimal layer

2.5 *Enrichment Analysis*

Enrichtment analysis adalah metode bioinformatika komputasional yang digunakan untuk mengevaluasi dan mengidentifikasi kelas gen atau protein yang direpresentasikan secara berlebihan (*over-represented*) di dalam suatu himpunan data molekuler berskala besar [60]. Pendekatan ini bekerja dengan memetakan daftar gen atau protein target ke dalam kamus dan basis data biologis yang terstandarisasi secara global, seperti *Gene Ontology* (GO) dan *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [10]. Secara spesifik, analisis GO berfungsi untuk mengklasifikasikan anotasi gen ke dalam tiga domain hierarkis utama, yaitu *Biological Process* (proses biologis molekuler), *Cellular Component* (lokasi anatomi seluler), dan *Molecular Function* (aktivitas biokimia) [27]. Sementara itu, analisis KEGG berfungsi untuk menempatkan dan memproyeksikan target-target genetik tersebut ke dalam konteks jejaring metabolik dan jalur persinyalan fungsional (*signaling pathways*) yang lebih luas [11].

Signifikansi utama dari analisis pengayaan terletak pada kemampuannya sebagai

jembatan yang menerjemahkan data statistik atau luaran matematis yang abstrak, seperti profil gen yang terekspresi secara diferensial maupun ekstraksi fitur dari algoritma *machine learning*, menjadi pemahaman mekanisme biologis yang nyata dan dapat diinterpretasikan [59]. Melalui analisis ini, peneliti dapat mengungkap rasionalitas mekanistik di balik fenomena polifarmakologi, di mana interaksi molekuler multitarget dapat secara bersamaan memodulasi homeostasis mikrolingkungan jaringan maupun regulasi sistem imun [18]. Tanpa tahapan pengayaan fungsional, luaran dari pemodelan komputasional tingkat tinggi hanya akan menjadi daftar identitas gen yang kaku, tanpa memberikan konteks fungsional yang esensial untuk penemuan terapeutik lanjutan [57].

2.6 Molecular Docking

Molecular docking merupakan metode komputasi berbasis struktur yang digunakan untuk memprediksi interaksi fisik antara ligan molekul kecil, seperti senyawa bioaktif herbal, dengan target makromolekuler spesifik pada kanker paru [24]. Tujuan utama metode ini adalah memprediksi orientasi dan konformasi optimal ligan pada sisi aktif (*active site*) protein target, serta memperkirakan afinitas ikatannya (*binding affinity*) [67]. Berbeda dengan pendekatan *machine learning* 1D atau 2D yang memanfaatkan representasi molekul sebagai deskriptor numerik abstrak, molecular docking memberikan interpretasi spasial 3D tingkat atomik yang memberikan pemahaman struktural mendalam untuk mengevaluasi interaksi protein-ligan secara sekuensial [26].

Secara mekanistik, proses ini melibatkan algoritma pencarian (*search algorithm*) untuk mengeksplorasi ruang konformasi (*conformational space*) ligan guna menghasilkan berbagai kandidat pose ikatan (*binding poses*) di dalam kantong reseptor [67]. Setiap pose tersebut kemudian dievaluasi menggunakan fungsi penilaian (*scoring function*) untuk mengestimasi nilai energi bebas ikatan (*binding free energy* atau ΔG) [67], [68]. Nilai ΔG total ini dihitung berdasarkan akumulasi komponen termodinamika yang mencakup gaya interaksi van der Waals, ikatan hidrogen, interaksi elektrostatik, serta energi sterik internal dari ligan itu sendiri [34].Nilai ΔG yang bernilai negatif menunjukkan bahwa pembentukan kompleks protein-ligan tersebut stabil secara termodinamika dan memiliki afinitas ikatan yang kuat [34]. Oleh karena itu, pendekatan ini menjadi langkah validasi krusial untuk mengonfirmasi kelayakan fisik dari senyawa bioaktif hasil prediksi model ansambel berbasis pohon sebelum dapat direkomendasikan lebih lanjut.