

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kanker merupakan salah satu penyebab utama kematian di seluruh dunia dengan beban penyakit yang terus meningkat seiring pertumbuhan populasi. Data GLOBOCAN 2020 menyoroti bahwa kanker payudara menempati urutan pertama kejadian kanker pada wanita, dengan satu dari setiap empat kasus kanker baru yang terdiagnosis secara global adalah kanker payudara (Sung *et al.*, 2021). Kanker payudara juga menjadi tantangan kesehatan masyarakat yang signifikan di tingkat nasional. Penyakit ini memiliki insiden serta mortalitas tertinggi di Indonesia. Laporan tahun 2022 mencatat angka kematian akibat kanker payudara mencapai 22.600 kasus (Rustamadji *et al.*, 2025). Tingginya angka kematian ini salah satunya disebabkan oleh heterogenitas respons pasien terhadap terapi standar. Pendekatan konvensional yang bersifat *one-size-fits-all* seringkali tidak efektif (Osei *et al.*, 2025). Kegagalan ini dikarenakan jarang memperhitungkan dasar molekuler kanker, yang merupakan penyakit genetik akibat mutasi, amplifikasi, atau ekspresi gen yang tidak normal. Hal ini mendorong perubahan paradigma menuju pengobatan presisi, di mana terapi disesuaikan dengan profil biologis spesifik dari setiap individu untuk meningkatkan efektivitas (Al-Matrafi *et al.*, 2025).

Pengembangan pengobatan presisi model komputasional yang mampu memprediksi efektivitas suatu obat berdasarkan karakteristik molekuler sel kanker. Analisis data omik atau data molekuler berskala besar dengan pendekatan *machine learning* telah muncul sebagai strategi yang menjanjikan untuk membangun model prediktif tersebut. Sebagian besar obat bekerja dengan menargetkan protein,

sehingga data proteomik memberikan representasi yang paling relevan terhadap keadaan sel. Hal ini diperkuat oleh studi komparatif berskala besar yang menunjukkan bahwa model berbasis data proteomik (protein) memiliki kekuatan prediktif yang lebih tinggi dibandingkan dengan model berbasis data transkriptomik (RNA) (Zheng *et al.*, 2024).

Performa model prediktif tetap sangat bergantung pada metode representasi numerik struktur kimia obat, terlepas dari ketersediaan basis data seluler yang optimal. Pendekatan klasik seperti *Morgan Fingerprint* banyak digunakan karena efisien, sebagaimana terlihat dalam studi terkait keunggulan data proteomik (Zheng *et al.*, 2024). Metode ini, bagaimanapun, memiliki keterbatasan mendasar akibat sifat representasinya bersifat tetap dan ketergantungan pada ada atau tidaknya substruktur kimia yang telah ditentukan sebelumnya. Risiko utama dari pendekatan tersebut adalah hilangnya informasi topologi molekuler yang bersifat kontekstual (Jiang *et al.*, 2021). Solusi atas keterbatasan tersebut ditawarkan oleh pendekatan *state-of-the-art* berbasis *Graph Neural Network* (GNN). GNN mampu mempelajari representasi fitur langsung dari struktur graf molekul sehingga dapat menangkap informasi kontekstual yang relevan dengan interaksi antara obat dan protein (Jiang *et al.*, 2021).

Penelitian ini berupaya mengisi celah tersebut dengan melakukan evaluasi sistematis terhadap berbagai arsitektur *machine learning* untuk prediksi respons obat pada kanker payudara berbasis data proteomik. Fokus utama studi terletak pada perbandingan dua metode representasi fitur obat, yaitu *Morgan Fingerprint* sebagai pendekatan klasik dan *Graph Neural Network* sebagai pendekatan modern. Kedua metode tersebut dikombinasikan dengan lima kelas algoritma prediktif yang

meliputi LightGBM, XGBoost, *Random Forest*, SVM, dan MLP. Evaluasi menyeluruh terhadap sepuluh kombinasi arsitektur yang dihasilkan diharapkan mampu mengidentifikasi model optimal. Model terpilih tersebut tidak hanya harus unggul dalam akurasi, tetapi juga memiliki tingkat *robustness* tinggi untuk mendukung pengembangan pengobatan presisi pada kanker payudara.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana perbandingan performa antara representasi fitur molekuler berbasis *Graph Neural Network* (GNN) dengan representasi berbasis *Morgan Fingerprint*?
2. Bagaimana perbandingan kinerja prediktif kelima model *machine learning* yang diuji (LightGBM, XGBoost, *Random Forest*, *Support Vector Machine*, *Multi-layer Perceptron*) serta model mana yang menunjukkan performa terbaik?
3. Bagaimana arsitektur *machine learning* paling optimal untuk memprediksi respons obat kanker payudara?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini bertujuan agar analisisnya lebih fokus dan tidak meluas dari pembahasan yang dimaksud. Oleh karena itu, penulis membatasi penelitian ini dalam ruang lingkup sebagai berikut:

1. Penelitian ini menggunakan data sekunder bersifat in-vitro dari *cell line* kanker yang bersumber dari basis data publik yaitu DepMap dan GDSC. Model yang

dihasilkan bersifat in-silico (komputasi) dan tidak divalidasi menggunakan data pasien klinis (in-vivo).

2. Ruang lingkup penelitian fokus pada kasus kanker payudara, sesuai dengan proses *filtering* data yang dilakukan menggunakan metadata *cell line*. Model tidak diuji pada jenis kanker lain.
3. Variabel yang digunakan sebagai prediktor terbatas pada data proteomik (protein). Penelitian ini tidak menggunakan data omik lain seperti genomik (DNA) atau transkriptomik (RNA).
4. Metode ekstraksi fitur obat yang dibandingkan terbatas pada dua representasi saja yaitu *Morgan Fingerprint* sebagai metode *baseline* klasik dan *Graph Neural Network* sebagai metode *state of the art*.
5. Evaluasi model prediksi terbatas pada lima model *machine learning* yang telah ditentukan, yaitu LightGBM, XGBoost, *Random Forest*, *Support Vector Machine*, dan *Multi-layer Perceptron*.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, tujuan dari penelitian ini sebagai berikut:

1. Membandingkan efektivitas metode ekstraksi fitur *Graph Neural Network* (GNN) dan *Morgan Fingerprint*.
2. Mengevaluasi dan membandingkan kinerja lima kelas model *machine learning* prediktif yang berbeda.
3. Mengidentifikasi satu arsitektur model (kombinasi fitur dan algoritma) yang paling optimal untuk memprediksi respons obat kanker payudara.